Mémoire présenté à l'Université Paris Diderot
pour l'obtention de l'habilitation à diriger des recherches

Spécialité : Mathématiques

Contributions à la théorie statistique de l'apprentissage :

sélection d'estimateurs et détection de ruptures

---

par

Sylvain Arlot

Chargé de recherches au CNRS, affecté à l'École normale supérieure

Soutenu publiquement le 3 décembre 2014 devant le jury composé de

Francis Bach	INRIA	Examinateur
Patrice Bertail	Université Paris Ouest Nanterre La Défense	Examinateur
Lucien Birgé	Université Pierre et Marie Curie	Président
Stéphane Boucheron	Université Paris Diderot	Rapporteur
Arnak Dalalyan	ENSAE	Rapporteur
Nicolas Vayatis	École Normale Supérieure de Cachan	Examinateur

et au vu du rapport également écrit par

Gábor Lugosi	ICREA / Universitat Pompeu Fabra	Rapporteur

Remerciements

Je souhaite tout d'abord remercier ici chaleureusement Stéphane Boucheron, Arnak Dalalyan et Gábor Lugosi d'avoir accepté de rapporter ce mémoire. Stéphane, tu as accepté d'être mon parrain à l'Université Paris Diderot, et de m'initier à certaines de ses arcanes administratives en ces temps troublés. Bien que je sache que « Someday, you'll call upon me to do a service for you », je voudrais te remercier particulièrement ici pour ton soutien. Arnak, nous nous croisons régulièrement depuis longtemps, à l'occasion de séminaires à Paris ou de conférences dans des lieux plus lointains, et j'ai toujours plaisir à discuter avec toi (de science ou d'autres choses) ; je suis particulièrement heureux que tu sois dans mon jury aujourd'hui en tant que rapporteur. Gábor, I am honored that you have accepted to review my habilitation thesis, and I am sorry that you finally were not able to come to my defense. We have only briefly met a few times up to now, and I hope that we will have opportunities for getting to know each other in the future.

Je remercie également Francis Bach, Patrice Bertail, Lucien Birgé et Nicolas Vayatis d'avoir accepté de faire partie de mon jury. Avoir l'occasion de soutenir mon habilitation devant vous est un honneur et sera (du moins, je l'espère) un plaisir.

J'ajouterais un mot pour Francis. Jean Ponce et toi m'avez accueilli à bras ouverts au sein de l'équipe Willow à mon arrivée à l'ENS. Tu m'as ensuite entraîné dans la création de l'équipe Sierra, et mis le pied à l'étrier en co-encadrant avec moi les thèses de Matthieu Solnon et Rémi Lajugie. C'est grâce à toi et à Jean que j'ai pu bénéficier d'un environnement de travail particulièrement fécond ; ce mémoire en présente une partie des fruits. Merci !

Je souhaite également remercier tout particulièrement Pascal Massart. Nous n'avons pas formellement collaboré depuis plusieurs années (encore que nous ayions un projet, qui je l'espère se concrétisera bientôt), mais tu es bien plus présent qu'il n'y paraît derrière tous les résultats présentés dans ce mémoire. Ton soutien sans faille et tes conseils toujours avisés me sont de précieux alliés depuis des années.

La recherche n'est pas un travail solitaire, et celle-ci passe aussi (d'abord ?) par l'enseignement. Je veux remercier ici Pierre Gaillard, Damien Garreau, Anisse Ismaili, Rémi Lajugie, Matthieu Solnon et Matteo Tanzi, que j'ai eu (et ai encore, pour Damien et Rémi) la chance d'encadrer (ou de co-encadrer) en master et/ou en thèse. Travailler avec vous m'a beaucoup appris, bien au-delà des quelques collaborations scientifiques qui apparaissent dans ce mémoire.

J'ai également eu la chance de collaborer avec quelques personnes sans lesquelles rien de ce qui suit n'aurait été possible : Alain Celisse, qui m'impressionne en arrivant toujours avant moi à mon bureau les matins où nous avons rendez-vous (en étant parti de Valenciennes !), Matthieu Lerasle, avec qui j'ai également eu le plaisir de partager une année de responsabilité du concours B/L de l'ENS (et qui a accepté de reprendre le flambeau, merci !), Peter Bartlett, qui m'a très gentiment accueilli à Berkeley pour deux courts séjours dont sont issus un article et une foule d'excellents souvenirs, ainsi que Gilles Blanchard, Robin Genuer, Zaïd Harchaoui, Nelo Magalhães, Étienne Roquain et Adrien Saumard. Je remercie également mes collaborateurs astronomes (Josselin Desmars, Valéry Lainey, Alain Vienne et mon père) et dynamiciens (Stefano Marmi et Duccio Papini), qui m'ont permis d'élargir mon horizon scientifique.

Je n'ai pas écrit d'article avec Aurélien Garivier et Gilles Stoltz, mais nous avons partagé bien plus encore. Aurélien, les trois années où nous avons partagé la responsabilité du concours B/L furent une formidable expérience humaine et mathématique. Tu nous as ensuite embarqués, Gilles et moi, dans l'aventure de la rédaction d'un livre d'exercices tirés de cette expérience. Je tiens à saluer ton efficacité, et ta patience face à nos maniaqueries typographico-linguistiques. Gilles, mon maître en typographie (j'espère que tu n'auras pas trop honte de ton élève en parcourant ce mémoire), je m'enorgueillis seulement de t'avoir appris quelques commandes LaTeX. Et bien au-delà de la typographie, j'ai énormément appris en travaillant avec toi.

Une composante importante de l'activité de recherche ne se matérialise pas sous forme de publications. Cette « partie immergée de l'iceberg » n'en est pas moins indispensable. Elle est faite de pauses café (ou thé), discussions informelles, séminaires, voyages, randonnées dans les calanques (et parfois tout ceci en même temps), bref, de bon moments partagés. Je tiens donc à remercier tous ceux que j'ai eu l'occasion de côtoyer professionnellement depuis une dizaine d'années, parmi lesquels je citerais mes collègues (anciens et actuels) des équipes Willow et Sierra, en particulier Guillaume Obozinski pour (entre autres) ces agréables moments que tu m'as permis de passer avec Compostelle, les membres du projet DETECT — Francis, Alain, Tristan, Josef, Étienne et Fanny —, Jean-Patrick Baudry, Yannig Goude et Olivier Wintenberger, pour de nombreuses discussions sur les séries temporelles qui n'ont (malheureusement) pas encore totalement porté leur fruits, ainsi que Nathalie Akakpo, Pierre Alquier, Yannick Baraud, Gérard Biau, Sébastien Bubeck, Olivier Catoni, Paul Doukhan, Christophe Giraud, Alexandre Gramfort, Magalie Fromont, Jonas Kahn, Béatrice Laurent, Guillaume Lecué, Erwan Le Pennec, Bertrand Michel, Pierre Neuvial, Patricia Reynaud-Bouret, Guillem Rigaill, Vincent Rivoirard, Jean-Philippe Vert et Nicolas Verzelen, pour n'en citer que quelques uns.

Merci également à tous ceux qui jouent un rôle de « support à la recherche » (sans qui le monde de la recherche s'écroulerait, donc) et que j'ai pu croiser lors de ces dix

dernières années. J'ai une pensée en particulier pour Pascal Chiettini, sans qui cette soutenance aurait probablement eu lieu en 2046.

Je remercie enfin mes amis, ma famille, Anne et Antoine, pour leur patience et leur soutien constants.

J'ajouterais un mot à l'intention de ceux que j'ai oublié de citer dans l'énumération ci-dessus, ou que je n'ai pas remercié autant que j'aurais dû. Je vous prie de bien vouloir m'en excuser : sachez que c'est à vous que je pense au moment où je relis ces remerciements.

Avant-propos

Ce mémoire présente, de manière synthétique, l'essentiel des travaux que j'ai effectués au cours de ma thèse (de septembre 2004 à décembre 2007) à l'Université Paris-Sud puis comme chargé de recherche au CNRS (depuis octobre 2008), au sein du Département d'Informatique de l'École Normale Supérieure.

Toutefois, mes travaux ne sont pas tous repris ici avec le même degré de précision. J'ai choisi de développer plus précisément certains d'entre eux, qui sont postérieurs à ma thèse et se situent au cœur de mon domaine de recherche : la construction de procédures de sélection d'estimateurs en apprentissage et leur étude sous l'angle de la statistique mathématique. J'ai également cherché à proposer (au chapitre 2) un point de vue plus général sur le problème de sélection d'estimateurs. Celui-ci reflète la démarche que j'ai adoptée dans mes travaux, mais aussi celle que bien d'autres ont utilisée avant moi, pour la sélection de modèles et au-delà.

Ce mémoire est composé de trois parties principales. Tout d'abord, le chapitre 1 est une présentation brève de mes travaux sur la sélection d'estimateurs, (presque) sans formule mathématique. Ensuite, les chapitres 2 à 6, rédigés en anglais, reviennent sur ces mêmes travaux de manière approfondie. Le premier de ces chapitres présente le cadre général utilisé dans ce mémoire, le problème de la sélection d'estimateurs, et une approche « générique » pour l'étudier. Les chapitres 3 à 6 (qui s'appuient tous sur le chapitre 2) sont largement indépendants entre eux. Ils reviennent successivement sur mes travaux autour de la validation croisée et des méthodes de rééchantillonnage (au chapitre 3), des méthodes de calibration par pénalités minimales (au chapitre 4), du problème de détection de ruptures (au chapitre 5), et de quelques autres problèmes de sélection d'estimateurs en apprentissage (au chapitre 6). Enfin, le chapitre 7, rédigé en anglais également, propose quelques perspectives dans le prolongement de ces travaux.

Productions scientifiques

Articles publiés dans des journaux

[1] Sylvain Arlot et Pascal Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10:245–279 (electronic), 2009.

[2] Josselin Desmars, Sylvain Arlot, Jean-Eudes Arlot, Valery Lainey et Alain Vienne. Estimating the accuracy of satellite ephemerides using the bootstrap method. *Astronomy and Astrophysics*, 499:321–330, Mai 2009.

[3] Sylvain Arlot. Model selection by resampling penalization. *Electronic Journal of Statistics*, 3:557–624 (electronic), 2009.

[4] Sylvain Arlot, Gilles Blanchard et Étienne Roquain. Some nonasymptotic results on resampling in high dimension, I: Confidence regions. *The Annals of Statistics*, 38(1):51–82, 2010.

[5] Sylvain Arlot, Gilles Blanchard et Étienne Roquain. Some nonasymptotic results on resampling in high dimension, II: Multiple tests. *The Annals of Statistics*, 38(1):83–99, 2010.

[6] Sylvain Arlot et Alain Celisse. Segmentation of the mean of heteroscedastic data via cross-validation. *Statistics and Computing*, pages 1–20, 2010.

[7] Sylvain Arlot et Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

[8] Sylvain Arlot et Peter L. Bartlett. Margin adaptive model selection in statistical learning. *Bernoulli*, 17(2):687–713, 2011.

[9] Matthieu Solnon, Sylvain Arlot et Francis Bach. Multi-task regression using minimal penalties. *Journal of Machine Learning Research*, 13:2773–2812 (electronic), Septembre 2012.

Articles publiés dans des actes de conférences très sélectives

[10] Sylvain Arlot, Gilles Blanchard et Étienne Roquain. Resampling-based confidence regions and multiple tests for a correlated random vector. In *Learning theory*, volume 4539 of *Lecture Notes in Computer Science*, pages 127–141. Springer, Berlin, 2007. COLT 2007.

[11] Sylvain Arlot et Francis Bach. Data-driven calibration of linear estimators with minimal penalties. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 46–54, 2009.

[12] Rémi Lajugie, Sylvain Arlot et Francis Bach. Large-margin metric learning for constrained partitioning problems. In *International Conference on Machine Learning (ICML)*, volume 32, pages 297–305, 2014. Voir aussi arXiv:1303.1280.

[13] Damien Garreau, Rémi Lajugie, Sylvain Arlot et Francis Bach. Metric learning for temporal sequence alignment. In *Advances in Neural Information Processing Systems 27*, 2014. Voir aussi arXiv:1409.3136.

L'article **[10]** est la version préliminaire des articles **[4]** et **[5]**.

Articles soumis pour publication

[14] Sylvain Arlot et Matthieu Lerasle. Why $V = 5$ is enough in V-fold cross-validation, 2014. arXiv:1210.5830v2.

Autres textes

Prépublications.

[15] Sylvain Arlot. Étude d'un modèle de dynamique des populations. Rapport technique 2004-38, Université Paris-Sud 11 Orsay, Septembre 2004. Mémoire de DEA. Directeur: Jean-Christophe Yoccoz. arXiv:1204.0799

[16] Sylvain Arlot. V-fold cross-validation improved: V-fold penalization, 2008. arXiv:0802.0566v2.

[17] Sylvain Arlot. Choosing a penalty for model selection in heteroscedastic regression, 2010. arXiv:0812.3141v2.

[18] Sylvain Arlot et Francis Bach. Data-driven calibration of linear estimators with minimal penalties, 2011. arXiv:0909.1884v2.

[19] Sylvain Arlot, Alain Celisse et Zaïd Harchaoui. Kernel change-point detection, 2012. arXiv:1202.3878v1.

[20] Sylvain Arlot et Robin Genuer. Analysis of purely random forests bias, 2014. arXiv:1407.3939v1.

[21] Rémi Lajugie, Piotr Bojanowski, Sylvain Arlot et Francis Bach. Semidefinite and spectral relaxations for multi-label classification, 2014. Preprint.

La prépublication **[18]** est une version longue de **[11]**.

Travaux en préparation.

[22] Sylvain Arlot. Minimal penalties and the slope heuristics: a survey, 2014.

Textes publiés (sans comité de lecture).

[23] Sylvain Arlot. V-fold penalization: an alternative to V-fold cross-validation. In *Oberwolfach Reports. Vol. 4, no. 4*, volume 4 of *Mathematisches Forschungsinstitut Oberwolfach Report*, pages 2951–2952. European Mathematical Society (EMS), Zürich, 2007. Report No.50/2007. Workshop: Reassessing the Paradigms of Statistical Model Building.

[24] Sylvain Arlot et Francis Bach. Data-driven penalties for linear estimators selection. In *Oberwolfach Reports. Vol. 7, no. 1*, volume 7 of *Mathematisches Forschungsinstitut Oberwolfach Report*. European Mathematical Society (EMS), Zürich, 2010. Report No.16/2010. Workshop: Modern Nonparametric Statistics: Going Beyond Asymptotic Minimax.

[25] Sylvain Arlot. Sélection de modèles. In *Journées MAS et Journée en l'honneur de Jacques Neveu*, Talence France, Août 2010. INRIA Bordeaux - Sud-Ouest.

[26] Rémi Lajugie, Francis Bach, and Sylvain Arlot. Metric learning for partitioning problems, December 2012. NIPS Workshop on Discrete Optimization in Machine Learning (DISCML) 2012. Disponible à l'adresse `http://las.ethz.ch/discml/discml12.html`

L'article **[23]** est une version courte de **[16]**, l'article **[24]** est une version courte de **[11]** et l'article **[26]** est une version courte de **[12]**. Le texte **[25]** est la présentation d'une session organisée au cours des Journées MAS 2010.

Thèse.

[27] Sylvain Arlot. *Rééchantillonnage et Sélection de modèles*. Thèse de doctorat, Université Paris-Sud 11, Décembre 2007. Disponible à l'adresse suivante : `http://tel.archives-ouvertes.fr/tel-00198803/`.

Les articles **[1]**, **[3]**, **[4]**, **[5]**, **[10]**, la prépublication **[16]** et une partie de la prépublication **[17]** sont issus de mon travail de thèse.

Codes informatiques

Les quatre codes informatiques suivants sont disponibles librement sur internet ; ils permettent de reproduire les expériences numériques réalisées dans les articles correspondants :

[28] Sylvain Arlot. Resampling Penalization for histogram selection in regression, 2009. Disponible à l'adresse `http://www.di.ens.fr/~arlot/code/RP.htm` Correspond à l'article **[3]**.

[29] Sylvain Arlot, Gilles Blanchard et Étienne Roquain. Confidence Regions and Multiple Tests by Resampling, 2009. Disponible à l'adresse `http://www.di.ens.fr/~arlot/code/CRMTR.htm` Correspond aux articles **[4]** et **[5]**.

[30] Sylvain Arlot et Alain Celisse. Change-Point Detection via Cross-Validation, 2010. Disponible à l'adresse http://www.di.ens.fr/~arlot/code/CHPTCV.htm Correspond à l'article **[6]**.

[31] Matthieu Solnon, Sylvain Arlot et Francis Bach. Boîte à outils pour la régression multitâches, 2012. Disponible à l'adresse http://www.di.ens.fr/~solnon/multitask_minpen_fr.html Correspond à l'article **[9]**.

Textes à vocation pédagogique

[32] Sylvain Arlot. Sélection de modèles et sélection d'estimateurs pour l'apprentissage statistique. Notes de cours (Cours Peccot, Collège de France), 2011. Disponible à l'adresse http://www.di.ens.fr/~arlot/peccot.htm

[33] Sylvain Arlot, Aurélien Garivier et Gilles Stoltz. *Exercices d'oral de mathématiques*. Collection Références sciences. Ellipses, 2015. Classes prépas BL - ECE - ECS. Corrigés et commentés par leurs auteurs. À paraître en mars 2015.

Table des matières

CHAPITRE 1

Résumé

Ce chapitre présente mes travaux de recherche d'une manière aussi peu technique que possible. Nécessairement, certains détails y sont laissés de côté. Un exposé technique détaillé (en anglais) est proposé aux chapitres 2 à 6. Nous encourageons le lecteur à s'y reporter aussi souvent qu'il le souhaite, de tels allers-retours étant facilités par le fait que ce chapitre suit le même plan que les chapitres 2 à 6.

1.1. Introduction

L'objectif principal de mes recherches est d'obtenir des résultats mathématiques pouvant aider les utilisateurs de méthodes d'apprentissage statistique, en particulier pour choisir leurs hyperparamètres ou pour choisir une méthode parmi plusieurs candidates. Ce but général comprend au moins quatre aspects importants.

Tout d'abord, certains algorithmes d'apprentissage sont connus pour fonctionner très bien en pratique, mais sans que des résultats théoriques ne l'expliquent vraiment. C'est par exemple le cas des forêts aléatoires [**Bre01**], un algorithme d'apprentissage très utilisé. Les résultats de **[20]** visent à comprendre les mécanismes qui font la réussite des forêts aléatoires, en analysant les « forêts purement aléatoires » qui ont été introduites car elles sont similaires aux forêts aléatoires de Breiman [**Bre01**] tout en étant plus simples à analyser théoriquement.

Par ailleurs, certaines méthodes d'apprentissage fonctionnent mieux que d'autres en pratique, alors que les garanties théoriques correspondantes ne reflètent pas du tout ces différences. Par exemple, parmi les méthodes de validation croisée **[7]**, la validation simple ou « hold-out » est peu utilisée car très instable en pratique, et on lui préfère en général la validation croisée par blocs « V-fold » avec $V = 5$ ou 10 [**BS92**, **HTF01**]. Pour autant, les garanties théoriques obtenues jusqu'à aujourd'hui sont aussi bonnes pour le hold-out que pour la validation croisée V-fold, voire légèrement meilleures pour le hold-out car celui-ci est beaucoup plus simple à analyser. Démontrer des résultats mathématiques qui reflètent ces différences empiriques est un problème important, et la section 1.4 présente quelques pas faits dans cette direction **[14]**.

Choisir V pour la validation croisée V-fold est un exemple au sein de la problématique plus générale de trouver des *compromis entre complexité algorithmique et performance statistique* en apprentissage automatique (« machine learning »). Lorsque plusieurs algorithmes d'apprentissage sont disponibles, avec une performance statistique

mais aussi un coût de calcul croissant, lequel est-il préférable de choisir à « budget » fixé (nombre d'observations, temps de calcul) ? Ce problème s'avère être l'une des questions clés induites par le défi d'analyser des quantités de données gigantesques [**BB08**], communément appelé « big data », et c'est récemment devenu un sujet de recherche actif [**CJ13**]. Dans le cas de la validation croisée V-fold, la complexité algorithmique est généralement proportionnelle à V, et quantifier précisément l'amélioration de la précision statistique obtenue quand V augmente est indispensable pour répondre à cette question du choix de V.

Un troisième point important est que l'analyse théorique des méthodes couramment utilisées permet non seulement de comprendre pourquoi elles fonctionnent, mais aussi de mettre en évidence leurs défauts et d'aider à les corriger. Ainsi, **[16]** démontre que la validation croisée V-fold estime le risque de manière biaisée, et que ce biais entraîne une performance sous-optimale pour la sélection d'estimateurs. En se fondant sur les pénalités par rééchantillonnage **[3]** déjà connues, il a été possible d'introduire les pénalités « V-fold » **[16]** qui corrigent le biais de la validation croisée V-fold sans modifier le temps de calcul. Les articles **[16]** et **[14]** ont alors pu démontrer l'optimalité des pénalités V-fold pour la sélection d'estimateurs, comme cela est détaillé en section 1.4.

Enfin, une approche théorique permet parfois de proposer de nouvelles méthodes d'apprentissage, qui répondent à des questions pratiques importantes et que l'on n'aurait sans doute pas pu construire d'une manière purement empirique. Par exemple, les méthodes dites de pénalisation dépendent souvent de constantes multiplicatives inconnues, telles que la variance du bruit en régression. En étudiant une question d'ordre théorique — quel est le niveau minimal de pénalisation nécessaire ? — Birgé et Massart [**BM07**] ont pu proposer une méthode (l'heuristique de pente) de choix automatique de la constante multiplicative optimale dans la pénalité. Cette méthode est décrite en section 1.5 et a démontré son intérêt pratique bien au-delà de son cadre initial [**BMM11**]. Les articles **[11]** et **[18]** proposent une généralisation de l'heuristique de pente, qui ne fonctionne qu'au prix d'une modification difficile à deviner empiriquement, mais qu'une analyse théorique suggère clairement. Le problème de détection de ruptures, décrit en section 1.6, fournit d'autres exemples d'algorithmes issus de travaux théoriques, dans deux cas où aucun algorithme n'était disponible auparavant. D'une part, un problème courant est de chercher à détecter des ruptures de moyenne dans une série temporelle sans que l'on puisse supposer la variance constante. Des considérations théoriques sur la régression hétéroscédastique **[16]**, **[17]** ont amené à modifier les techniques habituelles [**Leb05**] en proposant une nouvelle procédure fondée sur la validation croisée **[6]**. D'autre part, en utilisant des méthodes de sélection de modèles pour la détection de rupture dans le cadre des espaces de Hilbert à noyau reproduisant, une nouvelle procédure a pu être proposée dans **[19]**. Elle permet de traiter des cas que l'on retrouve dans diverses applications, où les observations sont multivariées, voire

à valeurs dans un ensemble qui n'est pas un espace vectoriel, tel que l'ensemble des séquences ADN ou l'ensemble des graphes finis.

Afin de fournir des réponses à ces quatre questions qui puissent être utiles en pratique, il faut *a minima* des résultats théoriques suffisamment précis pour être cohérents avec l'expérience. Dans l'idéal, il faudrait des résultats répondant à des questions pratiques pour lesquelles on ne dispose pas encore d'une réponse empirique claire. Ce minimum peut sembler bien peu, mais l'atteindre est déjà ambitieux pour de nombreux problèmes importants. Par exemple, démontrer théoriquement que la validation croisée V-fold est *strictement* meilleure que le hold-out pour la sélection d'estimateurs est un problème ouvert. Pour cette raison, nous faisons le choix de faire passer d'abord la précision des résultats obtenus, le niveau de généralité ne venant que dans un second temps. C'est ainsi que **[16]** ne considère « que » les régressogrammes, ou que **[14]** s'intéresse aux estimateurs par projection en estimation de densité avec le contraste des moindres carrés, qui n'est pas nécessairement le cadre le plus utilisé. Insistons sur le fait que disposer de résultats théoriques *précis* dans un cadre spécifique est utile pour le cas général, ne serait-ce que parce que cela permet de formuler des hypothèses précises qui peuvent être ensuite testées empiriquement dans d'autres cadres.

Tous les résultats présentés dans ce mémoire sont « non asymptotiques », c'est-à-dire qu'ils ne supposent pas que la taille de l'échantillon tend vers l'infini tandis que tous les autres paramètres (en particulier, la famille d'estimateurs considérés) restent constants. À l'inverse, les résultats asymptotiques cachent souvent dans des termes de reste de la forme $o(\cdot)$ des quantités qui peuvent s'avérer prédominantes en pratique, par exemple lorsque l'espace ambiant est de grande dimension. Cela ne veut pas dire que l'on cherche nécessairement des résultats s'appliquant à de très petits échantillons, ce qui serait difficile car les constantes numériques sont souvent surestimées dans les bornes théoriques non asymptotiques. L'intérêt principal d'une approche non asymptotique est que tous les paramètres du problème (dimension ambiante, rapport signal sur bruit, etc.) apparaissent explicitement dans les bornes obtenues, s'ils doivent y jouer un rôle, si bien que de tels résultats peuvent refléter fidèlement ce qui se produit en pratique.

1.2. Sélection d'estimateurs

Nous décrivons ici brièvement le cadre dans lequel se placent les résultats présentés dans ce mémoire.

Le problème général de l'inférence statistique (ou de l'apprentissage) peut se formuler ainsi : à partir d'un échantillon de n variables aléatoires indépendantes et de même loi P, on cherche à estimer une quantité $s^\star \in \mathbb{S}$ décrivant un aspect de cette loi. Pour cela, on se donne une fonction de perte $\mathcal{L} : \mathbb{S} \to \mathbb{R}$, minimale en $s^\star$ et s'écrivant

comme l'espérance sous P d'une fonction dite de « contraste ». Par exemple, un problème d'estimation de densité peut se formuler ainsi, $s^\star$ étant alors la densité de P par rapport à une mesure de référence.

L'apprentissage supervisé (ou « prédiction ») en est un autre exemple important : les observations sont des couples (X_i, Y_i) et l'on cherche à « prédire » la variable d'intérêt Y à partir de la seule connaissance de la variable explicative X. Ceci inclut notamment la régression (pour une variable d'intérêt continue) et la classification supervisée (pour une variable d'intérêt discrète).

Pour ce faire, on construit des estimateurs, qui associent à tout échantillon une estimation $\widehat{s} \in \mathbb{S}$ de la quantité cible $s^\star$. Un exemple important en statistique est celui des estimateurs par minimum de contraste : étant donné un « modèle », c'est-à-dire un sous-ensemble de $\mathbb{S}$, on choisit un élément $\widehat{s}$ du modèle qui minimise l'erreur commise sur les données (mesurée par un contraste empirique, aussi appelé risque empirique). L'idée est que si l'on cherche à minimiser la perte $\mathcal{L}$, qui est une espérance par rapport à P, en ne disposant pas de la loi P mais seulement d'un échantillon, on minimise à la place l'espérance de la même quantité relativement à la mesure empirique sur l'échantillon. Il est alors nécessaire de se restreindre à un modèle, car en s'autorisant $\mathbb{S}$ tout entier, on trouvera (presque) toujours un élément de $\mathbb{S}$ qui ne commet quasiment aucune erreur sur les n observations de l'échantillon, malgré le fait que les observations sont bruitées. Deux exemples classiques d'estimateurs par minimum de contraste sont les estimateurs du maximum de vraisemblance et les estimateurs des moindres carrés.

D'autres estimateurs sont bien sûr couramment utilisés en apprentissage statistique, comme par exemple les k plus proches voisins, la régression « ridge » à noyaux ou les forêts aléatoires. Nous renvoyons à [**DGL96**, **HTF01**, **GKKW02**, **BBL05**, **Was06**, **BvdG11**] pour compléter cette courte liste.

Le problème de sélection d'estimateurs est alors le suivant : étant donné une famille d'estimateurs, comment choisir en son sein (à l'aide des données uniquement) un estimateur dont la perte est aussi petite que possible ? Cette formulation générale recouvre au moins trois questions importantes en apprentissage statistique.

La sélection de modèles correspond au cas où l'on considère des estimateurs par minimum de contraste associés à différents modèles : sélectionner un estimateur revient alors à sélectionner l'un de ces modèles. Précisons qu'ici l'on ne cherche pas à identifier un « vrai » modèle (c'est-à-dire, qui contient la cible $s^\star$), mais plutôt à réaliser la meilleure estimation possible de $s^\star$, ce qui est un objectif différent [**Yan05**]. Pour des références bibliographiques sur la sélection de modèles, on pourra consulter [**BBM99**, **BA02**, **Mas07**, **HTF09**] par exemple.

Le choix d'hyperparamètres est une question omniprésente en apprentissage. On suppose qu'on a choisi une méthode d'apprentissage, par exemple, une méthode des plus proches voisins. Chaque méthode possède un ou plusieurs hyperparamètres (ici, le nombre de voisins considérés et une distance sur l'espace des variables explicatives) que

l'on ne peut pas déterminer *a priori* : leur valeur optimale dépend de caractéristiques de la loi P que l'on ne peut pas supposer connues à l'avance. Il faut donc utiliser les observations pour choisir ces hyperparamètres, ce qui constitue un problème de sélection d'estimateurs.

Enfin, une troisième situation — la plus générale — est celle où l'on envisage plusieurs méthodes, de différentes natures, pour analyser un même jeu de données, par exemple un estimateur des plus proches voisins, un estimateur par splines de lissage ou un estimateur paramétrique. Choisir parmi ces trois méthodes (et au passage choisir leurs hyperparamètres si nécessaire) est un problème de sélection d'estimateurs. Nous renvoyons à [**BGH10**, **Gir14**] pour des références bibliographiques sur la sélection d'estimateurs.

Le choix idéal — l'estimateur qui minimise la perte, appelé « oracle » — dépend de la loi P et est donc inconnu en pratique. Pour une méthode n'utilisant que les observations, l'objectif est donc de faire à peu près aussi bien que l'oracle. Plus précisément, si l'on définit la perte relative comme la différence entre la perte $\mathcal{L}$ et sa valeur minimale $\mathcal{L}(s^\star)$, on souhaite sélectionner un estimateur dont la perte relative est du même ordre de grandeur que la perte relative de l'oracle [1]. Une telle relation est appelée inégalité-oracle, et l'on parle d'inégalité-oracle « optimale » (au premier ordre) lorsque le rapport entre la perte relative de l'estimateur sélectionné et la perte relative de l'oracle tend vers 1 quand n tend vers l'infini.

Remarquons qu'au-delà du problème de sélection d'estimateurs, construire des procédures vérifiant une inégalité-oracle est également intéressant pour construire des estimateurs dits « adaptatifs au sens du minimax » [**BM97**, **BBM99**].

Quels sont les enjeux de la sélection d'estimateurs ? Considérons l'exemple de la sélection de modèles, où les modèles sont supposés être des espaces vectoriels de dimension finie emboîtés les uns dans les autres. D'un côté, avec un modèle de dimension très petite, on est certain de commettre une erreur du fait que la cible $s^\star$ n'appartient pas à ce modèle, voire en est assez éloigné : c'est l'*erreur d'approximation*. On parle de « sous-apprentissage », car il n'est pas possible dans un tel modèle d'apprendre suffisamment de paramètres pour bien estimer $s^\star$. À l'inverse, si l'on considère un modèle de dimension très grande (de l'ordre du nombre n d'observations), la cible $s^\star$ a toutes les chances d'en être proche et l'erreur d'approximation est donc quasi nulle. Cet avantage a un coût : un très grand nombre de paramètres doivent être estimés, à partir de données bruitées, d'où une deuxième source d'erreur, appelée *erreur d'estimation*. Au final, avec un grand

1. Il est ici important de ne pas considérer ici la perte mais la perte relative. Si on ne le fait pas, une garantie du type « la perte est inférieure à deux fois la perte de l'oracle » est très peu informative : même en supposant que l'oracle est consistant, c'est-à-dire que sa perte converge vers $\mathcal{L}(s^\star)$ lorsque n tend vers l'infini, cela n'implique pas la consistance de l'estimateur sélectionné, mais seulement une majoration (asymptotique) de sa perte par $2\mathcal{L}(s^\star) > \mathcal{L}(s^\star)$. C'est pourquoi, ici et aux chapitres suivants on considère presque toujours la perte relative $\ell(s^\star, \cdot)$ à la place de la perte $\mathcal{L}(t)$.

modèle, on obtient un estimateur qui « colle » aux observations (il y a suffisamment de paramètres dans le modèle pour reproduire assez précisément n'importe quel jeu de données bruité), mais qui ne s'approche pas de la cible $s^\star$: en faisant comme si les observations n'étaient pas bruitées, on ne peut pas généraliser convenablement. On parle alors de *sur-apprentissage* (« overfitting »). Pour résoudre le problème de sélection de modèles, il faut donc réaliser un compromis entre ces deux sources d'erreur, que l'on peut visualiser sur la figure 1 page 27.

Les mêmes phénomènes de sur-apprentissage et sous-apprentissage se produisent pour le problème général de sélection d'estimateurs comme illustré sur la figure 2 page 28, même si l'on ne peut pas toujours décomposer la perte relative en erreur d'approximation et erreur d'estimation. L'enjeu est donc identique : trouver un bon compromis entre ces deux situations extrêmes, ce que l'on formalise par une inégalité-oracle.

1.3. Une approche générale

Les procédures de sélection d'estimateurs les plus classiques sont définies comme la solution du problème de minimisation d'un critère $\mathcal{C}$ sur la famille d'estimateurs considérée. Si l'on note $\mathcal{R}$ la perte relative de ces estimateurs, notre objectif est donc de minimiser une quantité inconnue $\mathcal{R}$, et l'on minimise à la place une quantité connue $\mathcal{C}$ (fonction des observations seulement). C'est une technique souvent utilisée en apprentissage automatique ou en statistique, au-delà de la sélection d'estimateurs. Elle conduit par exemple aux estimateurs par minimum de contraste (on voudrait minimiser la perte, on minimise à la place un contraste empirique) ou à l'usage de relaxations convexes en optimisation (la quantité $\mathcal{R}$ est connue, mais la minimiser exactement est trop coûteux algorithmiquement, si bien qu'on résout à la place un problème de minimisation plus simple, par exemple en prenant pour $\mathcal{C}$ une majoration — ou relaxation — convexe de $\mathcal{R}$).

Pour construire un critère $\mathcal{C}$ adéquat, ou analyser une telle procédure étant donné $\mathcal{C}$, l'idée clé qu'utilisent un très grand nombre de travaux est un principe très simple, qui se résume formellement avec le lemme 2.1 en page 29, dont la démonstration est élémentaire. En quelques mots, si l'on dispose d'un encadrement pour la différence $\mathcal{C}-\mathcal{R}$ valable uniformément sur la famille d'estimateurs considérée, alors on a « presque » une inégalité-oracle.

Si de plus les deux bornes de cet encadrement sont négligeables devant $\mathcal{R}$ (uniformément sur la famille d'estimateurs considérée), alors on a une inégalité-oracle optimale au premier ordre. Ceci conduit au *principe d'estimation sans biais du risque*, sur lequel reposent notamment la validation croisée (présentée en section 1.4), le critère d'information d'Akaike [**Aka73**, AIC] et la pénalité C_p de Mallows [**Mal73**] qui est considérée en section 1.5 : utiliser un critère $\mathcal{C}$ dont l'espérance, pour chaque estimateur considéré individuellement, est égale à l'espérance de la perte relative $\mathcal{R}$ en cet estimateur. Le

lemme 2.1 valide ce principe, pourvu que l'on soit capables de démontrer une inégalité de concentration suffisamment précise pour $\mathcal{C} - \mathcal{R}$ autour de son espérance (zéro), uniformément sur la famille d'estimateurs.

On considère parfois de « grandes » familles d'estimateurs [**BM07**], pour lesquelles le principe d'estimation sans biais du risque ne fonctionne plus. C'est notamment le cas pour le problème de détection de ruptures, abordé en section 1.6. Ce dysfonctionnement est relié au fait que les déviations uniformes de $\mathcal{R}$ ou de $\mathcal{C}-\mathcal{R}$ sur la famille d'estimateurs sont alors supérieures d'un ordre de grandeur à l'espérance de $\mathcal{R}$. L'approche classique est d'utiliser encore une fois le lemme 2.1, mais en choisissant $\mathcal{C}$ de telle sorte que $\mathcal{C}-\mathcal{R} \geq 0$ pour tous les estimateurs. On en déduit alors une inégalité-oracle plus faible, garantissant que la perte relative de l'estimateur sélectionné est inférieure à la valeur minimale de $\mathcal{C}$. Si la majoration de $\mathcal{R}$ par $\mathcal{C}$ est suffisamment fine, la valeur minimale de $\mathcal{C}$ est (à peu près) de l'ordre de grandeur de la perte relative de l'oracle (la valeur minimale de $\mathcal{R}$).

Nous venons de décrire l'approche classique pour construire et analyser des procédures de sélection d'estimateurs. Pointons-en les limites.

Tout d'abord, comme cela a été mis en avant en introduction, un objectif important serait de pouvoir *comparer* des procédures de sélection d'estimateurs. Pour cela, le lemme 2.1 et les inégalités-oracle auxquelles il mène sont insuffisants : ce ne sont que des bornes supérieures sur la perte relative, et comparer de telles bornes est souvent trompeur. Au mieux, on peut espérer montrer qu'une procédure est optimale (au premier ordre), en démontrant qu'elle fait aussi bien que l'oracle (à des termes de reste près), puisque l'on sait qu'il est impossible de faire mieux que l'oracle. Si l'on veut réellement comparer deux procédures, il faut pouvoir démontrer une borne inférieure sur la perte relative de l'estimateur sélectionné par la première, et montrer que celle-ci est strictement supérieure à une borne supérieure sur la perte relative de l'estimateur sélectionné par la deuxième. Ceci nécessite, bien sûr, des bornes très précises.

L'approche usuelle pour démontrer des bornes inférieures est l'approche minimax : démontrer une borne inférieure en pire cas, comme par exemple dans l'article **[8]**. Ce type de résultat n'est pas toujours utile en pratique, car le pire cas est souvent très peu réaliste. C'est pourquoi, dans les articles **[16]** et **[17]**, des bornes inférieures sont prouvées en considérant des cas particuliers, certes, mais aussi « génériques » que possibles, c'est-à-dire, représentatifs de ce que l'on peut observer en pratique.

La deuxième limite importante de l'approche reposant sur le lemme 2.1 et le principe d'estimation sans biais du risque est qu'elles ne permettent pas vraiment de tenir compte de la « variance » du critère $\mathcal{C}$ dans l'analyse de ses performances moyennes pour la sélection d'estimateurs. Afin de combler ce manque, une heuristique est proposée dans l'article **[14]** et détaillée en section 2.2.5. En quelques mots, l'important n'est pas la variance de $\mathcal{C}$ pris en un estimateur (qui peut être modifiée sans changer la procédure de sélection d'estimateurs) mais plutôt la variance des incréments de $\mathcal{C}$, c'est-à-dire la

différence entre les valeurs qu'il prend en deux estimateurs quelconques. En particulier, il semble que si deux critères $\mathcal{C}_1$ et $\mathcal{C}_2$ ont la même espérance, mais que la variance des incréments de $\mathcal{C}_1$ est uniformément inférieure à celle des incréments de $\mathcal{C}_2$, alors la procédure de sélection d'estimateurs fondée sur $\mathcal{C}_1$ doit être meilleure que celle fondée sur $\mathcal{C}_2$. Cette heuristique est appliquée avec succès dans **[14]** pour l'analyse des performances de la validation croisée V-fold en fonction de V, comme expliqué en section 1.4.

1.4. Validation croisée et méthodes de rééchantillonnage

Comme on l'a vu en section 1.3, une approche naturelle pour construire un critère $\mathcal{C}$ est d'estimer (si possible, sans biais) la perte relative $\mathcal{R}$ de chacun des estimateurs de la famille considérée. Pour cela, idéalement, on aimerait pouvoir entraîner chaque estimateur sur les observations, puis le confronter à de *nouvelles* observations — indépendantes — afin d'évaluer ses performances. Ce n'est évidemment pas possible si l'on a déjà utilisé toutes les observations à notre disposition, et il n'est pas question d'utiliser deux fois les mêmes observations : cela conduirait au sur-apprentissage.

La validation croisée propose un moyen de faire « comme si » l'on avait de nouvelles observations, en découpant l'échantillon en deux : la première partie (l'échantillon d'entraînement) est utilisée pour entraîner chaque estimateur et la deuxième partie (l'échantillon de validation) est utilisée pour évaluer ses performances. Si l'on procède à un seul découpage, on parle de validation simple ou « hold-out ». Si l'on procède à plusieurs découpages et que l'on moyenne les évaluations des performances de l'estimateur considéré, on parle de validation croisée. L'usage est de fixer la valeur de la taille $n_t \in \{1, \dots, n-1\}$ de l'échantillon d'entraînement. Dans ce cas, si l'on considère tous les découpages possibles, on obtient le « leave-one-out » (« laisses-en un de côté ») lorsque $n_t = n-1$ et le « leave-p-out » en général (avec $p = n-n_t$). Souvent, pour réduire la complexité algorithmique, on ne considère qu'un petit nombre de découpages, la méthode la plus courante étant la validation croisée par blocs (« V-fold ») : on fixe une partition (régulière) de l'échantillon en V blocs, que l'on utilise successivement comme échantillon de validation (et le complémentaire comme échantillon d'entraînement). D'autres méthodes de validation croisée sont décrites dans l'article de survol **[7]**.

La validation croisée est ainsi un exemple d'application du principe de rééchantillonnage (ici, il s'agit même plus précisément de sous-échantillonnage), dont l'idée essentielle est de construire, à des fins statistiques, un ou plusieurs nouveaux échantillons à partir de l'unique jeu de données dont on dispose [**Efr79**].

La section 1.3 donne un angle d'attaque pour étudier les performances de la validation croisée pour la sélection d'estimateurs : calculer l'espérance du critère correspondant. En toute généralité, l'échantillon d'entraînement étant indépendant de l'échantillon de validation, l'espérance du critère par validation croisée est égale à l'espérance de la perte de l'estimateur considéré, *entraîné avec n_t observations* au lieu de n. Le calcul correspondant est détaillé en section 3.2, page 39. Le critère par validation croisée est

donc biaisé, ce biais étant (le plus souvent) faible si $n_t \sim n$ et d'autant plus fort que n_t est petit en comparaison de n.

En particulier, dans un cadre de régression, **[16]** démontre que la validation croisée V-fold est sous-optimale : si n est assez grand, avec grande probabilité, la perte relative de l'estimateur qu'elle sélectionne est supérieure ou égale à la perte relative de l'oracle multipliée par une constante $1+\kappa(V) > 1$. Malgré tout, une inégalité-oracle reste valable pour la validation croisée V-fold dans ce même cadre, mais elle n'est pas optimale : il y a un facteur multiplicatif $K(V) > 1$ devant la perte relative de l'oracle.

Comment obtenir une méthode optimale au premier ordre ? Soit l'on prend $n_t \sim n$ (par exemple le leave-one-out), mais avec les méthodes V-fold, cela nécessite de prendre V qui tend vers l'infini avec n, ce qui induit un temps de calcul souvent trop long. Une autre option, étudiée dans **[16]**, est de corriger le biais du critère par validation croisée.

L'idée est d'utiliser le principe de la validation croisée pour construire un critère pénalisé, égal à la somme du risque empirique et d'une « pénalité ». Cette pénalité vise à compenser l'optimisme du risque empirique comme estimateur de l'erreur commise sur de nouvelles observations. Efron [**Efr83**] a proposé d'utiliser le bootstrap pour construire une telle pénalité, et **[3]** démontre que celle-ci (parmi d'autres) fournit une procédure qui vérifie une inégalité-oracle optimale dans un cadre de régression.

En considérant le processus de sous-échantillonnage associé à la validation croisée V-fold, on obtient la méthode de « pénalisation V-fold », qui vérifie également une inégalité-oracle optimale dans un cadre de régression **[16]** et en estimation de densité **[14]**. Ces pénalités revisitent une méthode de correction du biais de la validation croisée proposée par Burman [**Bur89**]. Leur grand intérêt est d'avoir cette propriété d'optimalité (au premier ordre) tout en ayant une complexité algorithmique limitée, égale à celle des procédures par validation croisée V-fold habituellement utilisées.

Mentionnons pour finir que les articles **[16]** et **[14]** proposent deux approches pour obtenir de tels résultats, l'une ayant l'avantage d'être aisément généralisable (dès lors qu'une certaine inégalité de concentration peut être prouvée, voir la fin de la section 3.2), et l'autre d'être assez précise pour obtenir des bornes qui se comportent comme attendu en fonction de V.

À la suite de l'heuristique esquissée en section 1.3, comment prendre en compte la variance des critères par validation croisée dans cette analyse ? Les résultats mentionnés jusqu'à maintenant, comme la totalité des résultats sur les procédures de sélection d'estimateurs par validation croisée **[7]**, ne permettent en rien de distinguer le hold-out — critère réputé instable, dépendant du choix arbitraire d'un seul découpage et conduisant à de piètres performances pour la sélection d'estimateurs — des méthodes V-fold ou du leave-one-out, qui fournissent en pratique de bien meilleures performances en sélection d'estimateurs, apparemment parce qu'ils utilisent comme critère une moyenne sur plusieurs découpages.

Dans le cas des estimateurs par projection en estimation de densité **[14]**, la variance des incréments du critère par pénalisation V-fold peut être calculée exactement. Il s'avère qu'elle diminue avec V, conformément aux observations empiriques, approximativement comme

$$1 + \frac{4}{V-1} \ .$$

Ainsi, augmenter V (et donc le temps de calcul) doit améliorer les performances, mais cette amélioration se limite au gain d'une constante multiplicative dans la variance, et non pas d'un ordre de grandeur. Cette prévision, et plus généralement l'heuristique de la fin de la section 1.3, se trouvent confirmées par des expériences numériques **[14]**. Ceci explique donc en partie pourquoi il peut suffire de prendre $V = 5$ ou 10 pour la validation croisée V-fold, comme conseillé habituellement [**BS92**, **HTF01**] : avec $V = 10$, par exemple, la variance est quasiment identique à la valeur minimale possible (celle du leave-one out, c'est-à-dire $V = n$), avec une complexité algorithmique fortement réduite !

Mentionnons enfin d'autres travaux réalisés autour des méthodes par rééchantillonnage. D'une part, **[2]** valide empiriquement l'utilisation du bootstrap pour évaluer l'erreur d'extrapolation commise par des modèles dynamiques de la position de certains satellites de Saturne, de quelques dizaines à plusieurs centaines d'années hors de la période d'observation. Cette méthode a ensuite été employée pour d'autres applications en astronomie [**Des09**, **DAV10**]. D'autre part, des régions de confiance pour la moyenne de vecteurs gaussiens de grande dimension sont construites et validées dans **[10]** et **[4]**, ainsi que des procédures de tests multiples qui s'en déduisent **[5]**. Les expériences numériques réalisées dans ces trois articles indiquent que l'utilisation du rééchantillonnage dans ce contexte permet une adaptation automatique à la structure de corrélation des observations.

1.5. Pénalités minimales

Le principe de la pénalisation, évoqué à la section précédente, est de choisir l'estimateur qui minimise la somme du risque empirique et d'une « pénalité ». Cette dernière a pour rôle premier d'éviter le sur-apprentissage : si l'on minimisait uniquement le risque empirique, on sélectionnerait forcément un estimateur qui « colle » aux observations. Le critère que l'on cherche à minimiser étant la perte (relative), une pénalité idéale serait la différence entre la perte (relative) et le risque empirique, ou bien son espérance, en admettant le principe d'estimation sans biais du risque énoncé en section 1.3.

Dans le cas particulier de la régression par moindres carrés, en supposant le plan d'expérience (les X_i observés) fixe, on peut démontrer que l'espérance de cette pénalité idéale pour l'estimateur associé à un modèle de dimension D_m vaut

$$\frac{2\sigma^2 D_m}{n} \ ,$$

où σ^2 est la variance du bruit, supposée constante sur le plan d'expérience. Le calcul menant à cette pénalité, aussi appelée C_p [**Mal73**], est détaillé en section 4.3.

Cependant, en pratique, la variance σ^2 est inconnue et il est donc nécessaire de l'estimer, ce qui n'est pas un problème aisé **[22]**. Plus généralement, une pénalité optimale (ou quasi optimale) est souvent connue à une constante multiplicative près en pratique, pour plusieurs raisons.

(1) La pénalité optimale est connue théoriquement, mais dépend d'une quantité inconnue, par exemple σ^2 pour C_p et C_L en régression [**Mal73**].

(2) La pénalité optimale est connue théoriquement, mais seulement asymptotiquement. C'est le cas de AIC [**Aka73**] et BIC [**Sch78**] pour les méthodes du maximum de vraisemblance.

(3) La pénalité optimale peut être estimée par rééchantillonnage **[3]**, mais seulement à une constante multiplicative près, dont la valeur n'est pas toujours connue, ou alors seulement asymptotiquement.

(4) Une pénalité $C \times \mathrm{pen}_1$ satisfaisant une inégalité-oracle a constante multiplicative près est connue théoriquement, mais la valeur optimale de C n'est pas connue, ce qui se produit notamment pour la détection de ruptures [**CR04, Leb05**], qui est abordée en section 1.6, ou pour les complexités de Rademacher locales en classification [**BBM05, Kol06**].

Un remède à cette difficulté, l'heuristique de pente, a été proposé par Birgé et Massart [**BM07**] dans le cas de la régression par moindres carrés avec un plan d'expérience fixe. Le raisonnement est le suivant : que se passe-t-il si l'on utilise la pénalité

$$\frac{CD_m}{n}$$

pour une constante $C > 0$, en fonction de la valeur de C ? D'une part, si $C < \sigma^2$, on constate en calculant l'espérance du risque empirique que le critère pénalisé est une fonction décroissante de la dimension D_m des modèles : on sélectionne donc nécessairement l'un des plus gros modèles, il y a sur-apprentissage. D'autre part, si $C > \sigma^2$, le critère pénalisé devient strictement croissant au voisinage des plus gros modèles : on sélectionne donc un modèle de dimension bien plus petite, et l'on peut même démontrer une inégalité-oracle (sous-optimale si $C \neq 2\sigma^2$). Autrement dit,

$$\frac{\sigma^2 D_m}{n}$$

est une pénalité « minimale » pour ce problème de sélection de modèles, ce qui peut être prouvé théoriquement [**BM07**].

Le raisonnement que nous venons de décrire est purement théorique et pourrait sembler sans intérêt en pratique. Il n'en est rien, car on peut le compléter des deux remarques cruciales suivantes :

(i) La pénalité minimale est observable, car la dimension du modèle sélectionné avec la pénalité CD_m/n « saute » au voisinage de $C = \sigma^2$, et nulle part ailleurs, comme cela est représenté sur la figure 2 page 52.

(ii) La pénalité optimale est égale à deux fois la pénalité minimale.

Ces deux remarques fournissent un algorithme utilisable en pratique pour estimer la constante multiplicative optimale dans une pénalité : (i) déterminer la constante correspondant à la pénalité minimale, (ii) la multiplier par deux [**BM07**]. Une version de cet algorithme peut même être validée théoriquement, par une inégalité-oracle optimale au premier ordre valable sous des hypothèses minimales **[22]**, comme cela est expliqué en section 4.4.

Jusqu'où cette idée peut-elle être généralisée ? Des études empiriques [**BMM11**] ont montré que l'heuristique de pente fonctionne bien au-delà du cadre de [**BM07**]. Du point de vue théorique, l'article de survol **[22]** indique que des garanties ont été obtenues essentiellement pour des estimateurs des moindres carrés ou qui en sont « proches », en régression ou en estimation de densité [**Ler12**]. Il ne semble pas indispensable que les données soient indépendantes, au moins dans un cas [**Ler11**]. L'hypothèse d'avoir une variance du bruit constante en régression n'est pas non plus nécessaire, comme démontré par l'article **[1]**.

Toutefois, le fait que la pénalité optimale est égale à deux fois la pénalité minimale n'est pas valable en tout généralité. Ainsi, pour sélectionner parmi des estimateurs « linéaires » en régression (moindres carrés, plus proches voisins, Nadaraya-Watson, splines de lissage, etc.), **[11]** et **[18]** montrent que l'heuristique de pente, appliquée telle quelle, ne fonctionne pas. En revanche, en raisonnant sur les espérances du risque empirique et de la perte, **[11]** et **[18]** montrent que l'on peut encore utiliser la notion de pénalité minimale avec succès pour ce problème. La nouveauté essentielle est que les pénalités minimales et optimales ne sont pas proportionnelles, tout en étant chacune connue au facteur multiplicatif σ^2 près. Il en résulte un algorithme utilisable en pratique, pour lequel on peut obtenir une inégalité-oracle optimale au premier ordre **[11]**, **[18]**. Remarquons que l'on pourrait également penser à utiliser une méthode de validation croisée dans ce cadre. Elle satisferait certainement aussi une inégalité-oracle similaire (bien qu'aucun résultat de ce type n'ait été prouvé jusqu'à présent), mais l'expérience démontre qu'elles ont des performances inférieures ou égales à celles de la méthode de pénalisation proposée ci-dessus, et une complexité algorithmique bien supérieure. Mentionnons enfin que l'algorithme proposé par **[11]** et **[18]** est également utilisé pour un problème d'apprentissage « multitâches » **[9]**, comme expliqué en section 1.7.

1.6. Détection de ruptures

Le problème de détection de ruptures, aussi appelée segmentation unidimensionnelle, est classique en statistique [**BN93, BD93**]. Étant donné une série temporelle, dont la distribution change brusquement à certains instants inconnus (les « ruptures »),

l'objectif est de déterminer le nombre et la position de ces ruptures, comme illustré par la figure 1 page 63. Un tel problème est posé dans des domaines divers, par exemple pour l'analyse de signaux sonores [**HVLYFC09**], de données financières [**LT06**] ou biologiques [**Pic05**].

Les cas les plus classiques sont lorsque l'on cherche des ruptures dans la moyenne du signal (le reste de la distribution étant supposé inchangé) ou bien des ruptures indifféremment dues à un changement dans la moyenne ou dans la variance. Pour cela, une approche classique est de formuler ce problème comme un problème de sélection de modèles [**Yao88**, **YA89**, **CR04**, **Lav05**, **Leb05**, **BKL+09**].

Par exemple, la recherche de ruptures de moyenne peut être considérée comme un problème de régression, où chaque segmentation des données correspond à un modèle de fonctions constantes par morceaux. On se trouve alors dans le cas d'une grande famille de modèles, selon la terminologie employée en section 1.3, et l'approche générale de la section 1.3 conduit à des procédures de pénalisation pour la détection de ruptures qui vérifient des inégalités-oracle approchées [**CR04**, **Lav05**, **Leb05**]. De plus, le problème de minimisation correspondant peut être résolu efficacement par programmation dynamique, d'une manière exacte ou approchée [**Rig10**], ce qui n'est pas évident si l'on considère le problème « naïvement ».

D'autres problèmes de détection de rupture se posent cependant en pratique, pour lesquels on ne dispose pas toujours d'un algorithme approprié. Un premier exemple est celui de la recherche de ruptures dans la moyenne (uniquement) lorsqu'on sait que l'hypothèse d'une variance constante au cours du temps n'est pas vérifiée. C'est notamment le cas pour les données biologiques de type « CGH », pour des raisons liées au processus expérimental de mesure [**Pic05**].

Dans ce cas, des résultats théoriques sur la sélection de modèles en régression hétéroscédastique[2] indiquent — hors du cadre de la détection de rupture — que des pénalités du type de celles de [**Leb05**] ne sont pas adaptées à ce cadre **[17]**, au contraire des pénalités par rééchantillonnage **[3]** ou des pénalités V-fold **[16]**. L'échec des pénalités de [**Leb05**] dans ce cadre a pu être confirmé expérimentalement, conduisant à l'introduction d'une méthode de détection de ruptures fondée sur la validation croisée qui s'adapte effectivement à l'hétéroscédasticité des observations **[6]**. Il n'était pas évident qu'une telle méthode puisse être mise en œuvre avec une complexité algorithmique raisonnable. Ceci a été rendu possible par la combinaison de formules closes pour certains estimateurs par validation croisée [**Cel08**] avec un algorithme de programmation dynamique.

2. On parle de régression homoscédastique lorsque la variance du bruit est supposée constante, et à l'inverse de régression hétéroscédastique lorsque la variance du bruit est fonction de la variable explicative X.

Les méthodes de pénalisation évoquées ci-dessus [**CR04**, **Lav05**, **Leb05**] sont construites pour des données unidimensionnelles et peuvent être étendues directement au cas multivarié. Cependant, cette extension revient à considérer ces observations sous l'angle de la métrique euclidienne qui n'est pas toujours adaptée au problème considéré, tout particulièrement pour des données de grande dimension. De plus, certaines applications considèrent des données structurées complexes, telles que des histogrammes (pour l'analyse de données sonores ou vidéo), des chaînes de caractères (par exemple, des textes ou des séquences ADN) ou des graphes (en sociologie et en bioinformatique notamment).

Une procédure générale de détection de ruptures adaptée à de telles données est proposée par **[19]**. D'un point de vue abstrait, l'idée est de représenter les observations par des éléments d'un espace de Hilbert à noyau reproduisant, et d'appliquer à ces nouvelles « observations » une généralisation de l'approche de [**Leb05**].

Ceci est possible à mettre en œuvre avec une complexité algorithmique raisonnable grâce aux propriétés de ces espaces de Hilbert, en particulier « l'astuce du noyau » : l'algorithme de [**Leb05**] ne fait intervenir que des produits scalaires entre observations, sa version « à noyau » ne nécessite donc de calculer que les valeurs du noyau entre des paires d'observations. Ceci évite d'avoir à manipuler des éléments d'un espace de Hilbert qui est de dimension infinie pour la plupart des noyaux classiques.

Enfin, une inégalité-oracle dans **[19]** garantit que la procédure fonctionne — du moins si le noyau est bien choisi — et des expériences numériques montrent son intérêt pratique.

En particulier, une application un peu inattendue concerne les données unidimensionnelles : en utilisant la procédure de **[19]** avec un noyau bien choisi, on peut détecter des ruptures dans la distribution alors même que moyenne et variance restent constantes ! Dans un tel cadre, les procédures telles que celle de [**Leb05**] sont totalement inadaptées.

Pour des données multivariées de grande dimension, un défi pratique important est d'arriver à « apprendre » comment bien représenter ces données selon le problème de détection de ruptures que l'on cherche à résoudre. En particulier, il est très utile de pouvoir sélectionner les coordonnées informatives et éliminer celles qui ne le sont pas.

En supposant que l'on dispose d'exemples similaires de séries temporelles segmentées, l'article **[12]** propose une méthode algorithmiquement efficace pour « apprendre » ainsi une métrique appropriée à un problème de détection de ruptures donné. La difficulté principale est ici algorithmique : étant donné une fonction de contraste entre deux segmentations, il est facile d'en déduire un critère à minimiser pour apprendre la métrique, mais le problème d'optimisation correspondant n'est pas soluble exactement en un temps raisonnable. En s'inspirant d'une relaxation convexe proposée pour le problème de prédiction structurée [**THJA05**], l'article **[12]** résout ce problème de

minimisation d'une manière approchée, avec de bonnes performances expérimentales à la clé.

1.7. Autres problèmes d'apprentissage

En conclusion de ce chapitre, nous décrivons ci-dessous plusieurs autres travaux en lien avec le problème de sélection d'estimateurs pour différents problèmes d'apprentissage statistique.

Tout d'abord, la procédure d'apprentissage de métrique de **[12]** présentée à la fin de la section précédente s'étend à deux autres problèmes d'apprentissage non supervisé : la segmentation bidimensionnelle et la classification non supervisée. Une approche similaire est également proposée dans **[13]** pour le problème d'alignement dynamique de séquences, qui se pose par exemple en bioinformatique [**TPP99**] ou pour aligner deux enregistrements audio d'un même morceau de musique [**CSS+07**]. À chaque fois, le problème résolu relève principalement de l'optimisation (construire une relaxation adéquate) tout en ayant un but statistique : apprendre une bonne métrique pour le problème considéré.

Les forêts aléatoires [**Bre01**] sont très couramment utilisées en classification ou en régression, mais encore très mal comprises d'un point de vue théorique [**SBV14**]. Il s'agit d'une méthode d'ensemble : à partir d'un même échantillon, on ne construit pas un estimateur mais une grande famille d'estimateurs [3], chacun selon un processus aléatoire (en plus de l'aléa induit par l'échantillon). L'estimateur final est alors défini comme la moyenne (en régression) ou le résultat d'un vote majoritaire (en classification). L'heuristique sous-jacente est qu'en agrégeant des arbres de décision (les estimateurs) suffisamment différents les uns des autres, on améliore fortement la performance que l'on pourrait espérer avec un seul arbre de décision.

Une première étape vers l'étude des forêts aléatoires de Breiman [**Bre01**] est l'étude des forêts dites « purement aléatoires » [**Bre00**], qui sont un peu plus faciles à analyser théoriquement car elles supposent la structure (aléatoire) de chaque arbre indépendante de l'échantillon. Des garanties théoriques sont ainsi disponibles pour quelques forêts de ce type [**BDL08**, **Bia12**, **Gen12**]. Le risque des forêts « purement aléatoires » peut s'écrire comme la somme d'une erreur d'approximation et d'une erreur d'estimation, par analogie avec la décomposition décrite en section 1.2. Pour un exemple particulier de forêt, on sait que l'erreur d'estimation est inférieure d'un facteur 3/4 pour une forêt infinie, en comparaison d'un arbre pris isolément [**Gen12**]. Qu'en est-il pour l'erreur d'approximation ?

Dans le cas de la régression, l'article **[20]** démontre que pour trois exemples de forêts purement aléatoires au moins, une forêt infinie possède de meilleures propriétés

3. Chaque estimateur étant un arbre de décision, ils constituent ensemble une « forêt ».

d'approximation qu'un arbre. Plus précisément, en considérant une forêt infinie plutôt qu'un arbre, on gagne dans la *vitesse d'approximation* en fonction de la taille des arbres. Il en résulte que, si la taille des arbres est bien choisie, le risque d'une forêt (suffisamment grande) est inférieur d'un ordre de grandeur inférieur au risque d'un arbre seul. Ces résultats théoriques sont complétés par des observations expérimentales similaires pour un quatrième exemple de forêt purement aléatoire, qui est particulièrement proche des forêts aléatoires de Breiman [**Bre01**].

Les méthodes d'apprentissage dites « multitâches » sont populaires pour leur faculté à augmenter la taille « effective » d'un échantillon lorsqu'il est difficile voire impossible d'obtenir plus d'observations. L'idée est de considérer simultanément plusieurs problèmes similaires, avec l'espoir que cette similarité permettra de mieux résoudre chacun de ces problèmes que si on les avait considérés isolément. Par exemple, pour la classification d'images, il est naturel de penser que déterminer la présence d'un chat sur une image est une tâche similaire à la détermination de la présence d'un chien. La difficulté principale pour construire ou analyser de telles procédures est de bien identifier en quoi consiste cette similarité entre tâches et comment l'utiliser au mieux [**Sol13**].

Dans le cas de la régression, des estimateurs « ridge à noyaux » multitâches ont été proposés par [**EMP05**]. Ceux-ci dépendent d'un ou plusieurs paramètres, dont le choix constitue un problème de sélection d'estimateurs. Une approche par pénalisation est possible, mais la pénalité optimale qui résulte d'une analyse théorique du problème dépend de la matrice de covariance entre les différentes « tâches » **[9]**. En se fondant sur l'estimateur de la variance résiduelle issu de **[18]**, un estimateur de la matrice de covariance entre les tâches est proposé et validé théoriquement dans **[9]**. Il en résulte une inégalité-oracle optimale au premier ordre, et des simulations numériques démontrent que cette méthode peut effectivement obtenir de meilleures performances que si l'on considère les tâches séparément.

Un des défis majeurs de la sélection de modèle est l'adaptation à des propriétés inconnues (mais favorables pour l'apprentissage) des données que l'on cherche à analyser. Par exemple, en classification binaire, on sait que l'on peut obtenir de meilleures vitesses d'apprentissage lorsqu'une condition dite « de marge » [**MT99**] est vérifiée [**BBL05**, Section 5.5]. Intuitivement, cette condition donne une borne supérieure sur le niveau de bruit des observations, et plus cette borne est petite, meilleures sont les vitesses d'apprentissage que l'on peut espérer en pire cas.

À la suite de [**Kol06, MN06**], on peut remplacer cette condition par une condition plus faible et « locale », c'est-à-dire ne dépendant que de propriétés de la loi P des observations « à l'intérieur » du modèle considéré. La question qui se pose alors est de savoir s'il est possible de construire une procédure de sélection de modèles qui s'adapte à cette condition de marge « locale », c'est-à-dire, qui fasse (à constante près) aussi bien que l'estimateur de minimisation du risque empirique sur le « meilleur » modèle. L'étude de cette question est l'objet de l'article **[8]** où sont démontrés deux résultats principaux.

D'une part, pour une famille de modèles emboîtés, l'adaptation à la condition de marge locale est possible ; on l'obtient par exemple avec une procédure de pénalisation par une complexité de Rademacher locale [**BMP04**, **LW04**, **BBM05**, **Kol06**]. D'autre part, un contre-exemple montre qu'une telle adaptation est impossible en toute généralité.

La classification multiétiquettes correspond au cas où la variable d'intérêt Y est un sous-ensemble d'un ensemble $\mathcal{V}$ d'étiquettes possibles. Par exemple, une image, une vidéo [**XHE+10**] ou un texte [**Joa98**] peuvent être annotés avec plusieurs étiquettes simultanément. La difficulté est alors que l'ensemble des possibles pour Y est en général gigantesque, de taille $2^{\mathrm{Card}(\mathcal{V})}$, bien plus grand que le nombre d'observations, voire trop grand pour être parcouru ne serait-ce qu'une fois.

Une approche élémentaire est de considérer les $\mathrm{Card}(\mathcal{V})$ problèmes de classification binaire associés à chacune des étiquettes séparément. C'est l'approche « one versus rest » (OvR). Son défaut est qu'elle ne tient pas compte des relations possibles entre étiquettes. Par exemple, on a plus de chances de voir sur la même image un zèbre et un lion que de voir ensemble un zèbre et un caribou. Une approche « multitâches » capable de prendre en compte ces relations a donc toutes les chances de fonctionner mieux que OvR.

Une manière d'encoder des corrélations positives ou négatives entre étiquettes est proposée par **[21]**, au sein d'un algorithme qui peut être mis en œuvre pour des ensembles $\mathcal{V}$ de taille relativement grande (jusqu'à 159 étiquettes possibles dans les expériences numériques de **[21]**). Celle-ci repose sur une relaxation convexe issue de [**THJA05**], le problème de classification multiétiquettes étant un problème de prédiction structurée. En comparaison de l'état de l'art sur la classification multiétiquettes, une nouveauté importante de **[21]** est la possibilité d'apprendre à la fois des corrélations positives et négatives. Des expériences sur données réelles montrent qu'il en résulte une amélioration des performances de classification.

CHAPTER 2

Estimator selection: a general point of view

The main concern of my work is to provide mathematical results that can help practitioners choose or tune a learning method in order to analyze some data. This general goal includes four important aspects.

First, some learning methods are empirically known to work well, but few theoretical results are available for explaining their good performance. This is for instance the case with random forests [**Bre01**], a very popular learning method. Among others, the results of **[20]** contribute to understanding why random forests work so well, by analyzing "purely random forests", that have been introduced because they are similar to the original algorithm of [**Bre01**] while being easier to analyze theoretically; these results are presented in Section 6.2.

Second, some learning methods empirically perform much better than others, while theoretical guarantees do not at all account for such differences. For instance, among cross-validation methods **[7]**, hold-out is known to work poorly compared to V-fold cross-validation with $V \geq 5$, and a classical advice [**BS92, HTF01**] is to take V between 5 and 10. Nevertheless, up to now, theoretical guarantees are as good for hold-out as for V-fold cross-validation, and often slightly better for hold-out, because hold-out is much easier to analyze, see for instance [**BM06**]. Providing theoretical results that reflect these empirical differences is an important problem, and some results in this direction have been obtained in **[14]**; they are presented in Chapter 3.

The question of choosing V for V-fold cross-validation methods is related to the more general challenge of solving *computational trade-offs* in machine learning. When several learning algorithms are available, with increasing computational complexity and statistical performance, which one should be used, given the amount of data and the computational power available? This problem has emerged as a key question induced by the challenge of analyzing large amounts of data [**BB08**]—the "big data" challenge—and it has been tackled in various settings in the recent years [**CJ13**]. In the case of V-fold cross-validation, the computational complexity is usually proportional to V, and quantifying the increase of the statistical performance as a function of V is crucial for answering this question.

Third, a theoretical analysis of existing methods is not only useful for explaining why they work, but can also point out their drawbacks and help correct them. For instance, **[16]** shows that V-fold cross-validation is biased for risk estimation, and that this bias

can induce suboptimal performance for estimator selection. Then, V-fold penalization—which is closely related to bias-corrected V-fold cross-validation [**Bur89**]—has been introduced as a natural way to correct for this bias, and it has been proved to be optimal for estimator selection in **[16]** and in **[14]**, see Chapter 3.

Fourth, new learning algorithms can emerge from theoretical considerations, thus providing an answer to some important practical questions. For instance, penalization methods often depend on unknown multiplicative constants, such as the residual variance in regression. The theoretical concept of minimal penalty, studied by [**BM07**], lead to designing a method—called the slope heuristics—for estimating the optimal multiplicative constant in the penalty from data only. This method, presented in Chapter 4, has been proved useful in practice in various settings [**BMM11**], far beyond the framework of [**BM07**]. In **[11]** and **[18]**, the slope heuristics of [**BM07**] is generalized to a wider setting. Notably, this generalization is not straightforward and could not be guessed exclusively from numerical experiments: a theoretical approach was necessary to derive it. Other examples of algorithms coming from theoretical works can be found for the change-point detection problem, described in Chapter 5, in two situations where no algorithm was available previously. When the goal is to detect changes in the mean of some time series while the variance of data might not be constant—a common practical situation—, a new cross-validation based algorithm is proposed in **[6]**, which has its roots in theoretical considerations on regressograms and cross-validation methods in heteroscedastic regression, namely in **[16]** and **[17]**. When the time series to be analyzed is multivariate, or takes its values in a general set $\mathcal{X}$—not necessarily a vector space, for instance the set of DNA sequences or of finite connected graphs—, a new algorithm is proposed and analyzed in **[19]**. In short, it comes from applying the model selection approach to change-point detection in the context of reproducing kernel Hilbert spaces.

In order to provide an answer to one of these four questions that can be useful to practitioners, theoretical results must at least be precise enough to be consistent with what is widely known empirically. Ideally, they should solve practical issues that have no clear empirical answer yet. The above minimal requirement might seem very low, but it is actually challenging for many important problems. For instance, showing theoretically that V-fold cross-validation works strictly better than hold-out for estimator selection is still a difficult open problem. To this aim, precision of the theoretical results comes first, sometimes requiring as a first step to decrease the level of generality—hence considering regressograms in **[16]**—, or to focus on frameworks and estimators that are not the most widely used—for instance least-squares density estimation and projection estimators in **[14]**. Nevertheless, precise theoretical results in a specific setting are useful for the general case, in particular because they lead to formulating precise hypotheses that can be tested empirically in other settings.

All results presented here are non-asymptotic, that is, we do not assume that the sample size tends to infinity while all other parameters of the problem stay fixed. On

the contrary, asymptotic results often hide inside $o(\cdot)$ some remainder terms that can be dominating in practice, in particular when considering high-dimensional data. This does not mean that the goal is to deal with very small sample sizes, in particular because numerical constants often are pessimistic in non-asymptotic results. The main advantage of the non-asymptotic approach is that all parameters appear explicitly in the bounds if necessary, in particular the ambient dimension and the signal-to-noise ratio, so that such results can reflect faithfully what happens in practice.

In the following we mostly consider the estimator selection problem, which is precisely defined in Section 2.1. Estimator selection includes several important problems of statistical learning—model selection, hyperparameter tuning, data-driven choice among learning algorithms of different nature—and is a fruitful approach to deal with specific but important issues such as change-point detection. A general approach to estimator selection is sketched in Section 2.2, then applied to three different contexts: cross-validation and resampling methods in Chapter 3, minimal penalties for calibrating multiplicative constants in penalization methods in Chapter 4, and change-point detection seen as a model selection problem in Chapter 5. Other works on estimator selection (or related issues) for various statistical learning problems are described in Chapter 6.

2.1. The estimator selection problem

This section formally introduces the estimator selection problem and the framework we consider here and in the following chapters. A more detailed account on estimator selection, in particular model selection, can be found in Sections 1–3 of **[7]** and in [**Mas07**] for instance.

2.1.1. The supervised learning framework. The main setting we have in mind is supervised learning, also known as the prediction problem. Assume that we observe

$$(X_1, Y_1), \dots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y},$$

independent random variables with (unknown) common distribution P. Given a "new observation" (X_{n+1}, Y_{n+1}), that is, a random variable with distribution P and independent from the sample $D_n = (X_i, Y_i)_{1 \leq i \leq n}$, the goal is to be able to predict the value of Y_{n+1} (the variable of interest) when only X_{n+1} (the explanatory variable) is observed. Formally, we want to build from the sample D_n a predictor t, that is, a measurable mapping $\mathcal{X} \to \mathcal{Y}$ such that $t(X_{n+1})$ "predicts" Y_{n+1}. Then, we want to minimize the average accuracy of t, that is, its loss

$$\mathcal{L}(t) := \mathbb{E}_{(X_{n+1}, Y_{n+1}) \sim P}\Big[d\big(t(X_{n+1}), Y_{n+1}\big)\Big]$$

for some given function $d : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

2.1.2. General framework. We actually consider in the following a slightly more general framework, that allows in particular to simultaneously consider supervised learning and density estimation in Section 2.2 and in Chapter 3.

Let Ξ be some measurable space, P some (unknown) distribution on Ξ and assume that we observe $\xi_1, \dots, \xi_n \in \Xi$ independent random variables with common distribution P. The purpose of statistical inference is to estimate from the data $D_n = (\xi_i)_{1 \le i \le n}$ some target feature $s^\star$ of the unknown distribution P—for instance, the density of P with respect to some reference measure μ on Ξ, or the regression function.

Let $\mathbb{S}$ denote the set of possible values for $s^\star$. The quality of $t \in \mathbb{S}$, as an approximation to $s^\star$, is measured by its loss $\mathcal{L}(t)$ where $\mathcal{L} : \mathbb{S} \to \mathbb{R}$ is called the *loss function*; the loss is assumed to be minimal for $t = s^\star$. In the following we consider loss functions that can be defined by

$$\forall t \in \mathbb{S}, \quad \mathcal{L}(t) := \mathbb{E}_{\xi \sim P}\big[\gamma(t;\xi)\big] = P\gamma(t) \ , \tag{2.1}$$

where $\gamma : \mathbb{S} \times \Xi \to \mathbb{R}$ is called a *contrast function*. In Eq. (2.1), the notation $P\gamma(t)$ means that the function $\xi \mapsto \gamma(t;\xi)$ is integrated with respect to the measure P on Ξ. For $t \in \mathbb{S}$, the quantity $P\gamma(t)$ measures the average discrepancy between t and a new observation ξ with distribution P.

Given a loss function $\mathcal{L}(\cdot)$, two useful quantities are the *excess loss*

$$\ell(s^\star, t) := \mathcal{L}(t) - \mathcal{L}(s^\star) \ge 0$$

and the *risk of an estimator* $\widehat{s}(\xi_1, \dots, \xi_n)$ of the target $s^\star$, which is defined as

$$\mathbb{E}_{\xi_1,\dots,\xi_n \sim P}\Big[\ell\big(s^\star, \widehat{s}(\xi_1, \dots, \xi_n)\big)\Big] \ .$$

2.1.3. Examples. The following four important statistical learning problems can be formulated as examples of the general framework of Section 2.1.2.

Example 2.1 (Supervised learning or prediction)**.** If $\Xi = \mathcal{X} \times \mathcal{Y}$, $\mathbb{S}$ is the set of measurable mappings $\mathcal{X} \to \mathcal{Y}$ (predictors) and for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $t \in \mathbb{S}$,

$$\gamma\big(t;(x,y)\big) = d\big(t(x), y\big) \ ,$$

we recover the supervised learning problem, as described in Section 2.1.1. Then, the target $s^\star$ is any element of $\operatorname{argmin}_{t \in \mathbb{S}} \mathcal{L}(t)$—assuming the argmin is not empty.

Two important supervised learning problems are regression (when $\mathcal{Y}$ is continuous) and classification (when $\mathcal{Y}$ is discrete).

Example 2.2 (Regression)**.** When $\mathcal{Y} = \mathbb{R}$ (or $\mathbb{R}^k$ for multivariate regression), Example 2.1 is the regression problem. Generally, the feature space $\mathcal{X}$ is a subset of $\mathbb{R}^\ell$. Let η denote the regression function, that is, $\eta(X) = \mathbb{E}_{(X,Y)\sim P}[Y \,|\, X]$. Then,

$$\forall i \in \{1, \dots, n\}, \quad Y_i = \eta(X_i) + \varepsilon_i \qquad \text{with} \quad \mathbb{E}[\varepsilon_i \,|\, X_i] = 0 \ .$$

A popular contrast in regression is the *least-squares contrast*

$$\gamma\big(t;(x,y)\big) := \big(t(x) - y\big)^2 ,$$

for which $P\gamma(t)$ is minimal over $\mathbb{S}$ for $t = \eta = s^\star$, and the excess loss is

$$\ell(s^\star, t) = \mathbb{E}_{(X,Y)\sim P}\Big[\big(s^\star(X) - t(X)\big)^2\Big] \ .$$

Note that the excess loss of t is the square of the L^2 distance between t and $s^\star$, so that prediction and estimation here are equivalent goals.

Example 2.3 (Classification)**.** When $\mathcal{Y}$ is finite, Example 2.1 is the (supervised) classification problem. In particular, $\mathcal{Y} = \{0, 1\}$ corresponds to *binary* (*supervised*) *classification*.

With the 0–1 contrast function $\gamma\big(t;(x,y)\big) = \mathbb{1}_{t(x)\neq y}$, the minimizer of the loss is the so-called Bayes classifier $s^\star$ defined by

$$\forall x \in \mathcal{X}, \quad s^\star(x) = \mathbb{1}_{\eta(x)\geq 1/2} \ ,$$

where η denotes the regression function $\eta(X) = \mathbb{P}_{(X,Y)\sim P}(Y = 1 \,|\, X)$.

Some learning problems can also be cast into the general framework of Section 2.1.2, without being instances of Example 2.1, for instance, density estimation, described in Example 2.4 below, and classification with convex losses, see [**BBL05**].

Example 2.4 (Density estimation)**.** If μ is some reference measure on Ξ that dominates P, $s^\star$ the density of P with respect to μ and $\mathbb{S}$ the set of densities on Ξ with respect to μ, then the general framework reduces to density estimation.

At least two losses of the form of Eq. (2.1) are minimal over $\mathbb{S}$ at $s^\star$. First, taking $\gamma(t;x) = \|t\|^2_{L^2(\mu)} - 2t(x)$ the least-squares contrast, the excess loss

$$\ell(s^\star, t) = \|t - s^\star\|^2_{L^2(\mu)}$$

is the L^2 distance between densities t and $s^\star$. Second, taking $\gamma(t;x) = -\log\big(t(x)\big)$ the log-likelihood contrast, the excess loss

$$\ell(s^\star, t) = \mathbb{E}_{\xi\sim P}\left[\log\left(\frac{s^\star(\xi)}{t(\xi)}\right)\right] = \int s^\star \log\left(\frac{s^\star}{t}\right)\mathrm{d}\mu$$

is the Kullback-Leibler divergence between distributions $t\mu$ and $s^\star\mu$.

2.1.4. Estimators or statistical algorithms. Let us call *estimator* or *statistical algorithm* any measurable mapping $\widehat{s} : \bigcup_{n\in\mathbb{N}} \Xi^n \to \mathbb{S}$. For any $D_n = (\xi_i)_{1\leq i\leq n} \in \Xi^n$—that we call a sample—the output of $\widehat{s}$, denoted by $\widehat{s}(D_n) \in \mathbb{S}$, is an estimator of $s^\star$. The quality of $\widehat{s}$ is then measured by $\mathcal{L}(\widehat{s}(D_n))$ or its expectation, which should be as small as possible. Note that here and in the following, both elements of $\mathbb{S}$ and mappings $\bigcup_{n\in\mathbb{N}} \Xi^n \to \mathbb{S}$ are called estimators. Such an abuse of language is usual in the learning literature, so that we only use the (less common) term "statistical algorithm" for cases

where there might be some confusion. Similarly, as usual in statistics, we often write $\widehat{s}$ as a shortcut for $\widehat{s}(D_n)$, when no confusion is possible.

Many estimators have been proposed in statistics and learning, and we do not try here to list even the most classical ones. We only mention a few examples of particular interest in the following and we refer to [**DGL96**, **HTF01**, **GKKW02**, **BBL05**, **Was06**, **BvdG11**] for other classical examples.

Minimum contrast estimators refer to a classical family of statistical algorithms. Given some subset S of $\mathbb{S}$ called a *model*, a minimum contrast estimator over S is any $\widehat{s}(D_n) \in \mathbb{S}$ that minimizes over S the empirical contrast

$$t \mapsto P_n\gamma(t) = \frac{1}{n}\sum_{i=1}^{n} \gamma(t;\xi_i) \qquad \text{where} \quad P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\xi_i} \ .$$

The idea is that the empirical contrast $P_n\gamma(t)$ has an expectation $P\gamma(t)$ which is minimal for $t = s^\star$. Minimizing $P_n\gamma(t)$ over a set S of candidate values for $s^\star$ hopefully leads to a good estimator of $s^\star$.

In supervised learning (Example 2.1), minimum contrast estimators are often called empirical risk minimizers [**Vap82**], in particular (but not only) in binary classification with the 0–1 constrast.

In regression (Example 2.2), taking the least-squares contrast

$$\gamma\big(t;(x,y)\big) = \big(t(x) - y\big)^2$$

leads to least-squares (or projection) estimators. If in addition S is the set of piecewise constant functions on some fixed partition of $\mathcal{X}$, we get a regressogram estimator.

In density estimation (Example 2.4), taking the least-squares contrast

$$\gamma(t;\xi) = \|t\|^2_{L^2(\mu)} - 2t(\xi)$$

leads to projection estimators, and taking the log-likelihood contrast

$$\gamma(t;x) = -\log\big(t(x)\big)$$

leads to maximum-likelihood estimators.

Among other estimators considered in the following, let us also mention:

- local averaging estimators for regression or classification [**DGL96**], such as Nadaraya-Watson kernel estimators [**Nad64**, **Wat64**] and k-nearest neighbors [**FH51**, **FH89**, **CH06**],
- (kernel) ridge regression [**SS01**] and spline smoothing [**Wah90**],
- classification and regression trees [**BFOS84**] and random forests [**Bre01**].

2.1.5. Estimator selection. Assume that a finite or countable family $(\widehat{s}_m)_{m\in\mathcal{M}}$ of statistical algorithms (estimators) and a loss function $\mathcal{L}$ are given. Then, the estimator selection problem is to choose among

$$\big(\widehat{s}_m(D_n)\big)_{m\in\mathcal{M}} ,$$

that is, to choose some $\widehat{m}(D_n) \in \mathcal{M}$ such that

$$\ell\big(s^\star, \widehat{s}_{\widehat{m}(D_n)}(D_n)\big)$$

is as small as possible. This formulation of the problem is often called "estimator selection for estimation", because other goals can be pursued in some specific situations, such as trying to identify the smallest correct model in model selection, see [**Yan05**] and Sections 2.3–2.4 in [7]. In the following, we focus on estimator selection for estimation. Note also that our framework is non-asymptotic, so we allow the family $(\widehat{s}_m)_{m\in\mathcal{M}}$ to vary with n: when more data are available, one can reasonably consider more estimators, for instance by taking into account more features. This possible dependence on n sometimes needs to be made explicit, in which case we write $\mathcal{M}_n$ instead of $\mathcal{M}$.

The estimator selection problem includes at least three important challenges of statistical learning. *Model selection* corresponds to the case where a contrast function γ and a family $(S_m)_{m\in\mathcal{M}}$ of models is given, and the goal is to select among the corresponding minimum contrast estimators $\widehat{s}_m$, as defined in Section 2.1.4. Another situation is when a learning method has already been chosen—say, k-nearest neighbors—and it remains to choose its hyperparameters—here, the number k of neighbors and maybe also a distance d on $\mathcal{X}$. Then, if $\widehat{s}_m$ denotes the corresponding learning method using a fixed set m of hyperparameters, the estimator selection problem corresponds to *hyperparameter tuning*. Finally, when several learning methods of different nature are considered for analyzing a given data set, for instance k-nearest neighbors, spline smoothing and some parametric estimator, choosing among these different methods also is an estimator selection problem.

Since the goal is to minimize the excess loss of the final estimator $\widehat{s}_{\widehat{m}}$, the best possible choice is the so-called *oracle* estimator $\widehat{s}_{m^\star}$ where

$$m^\star = m^\star(D_n) \in \operatorname*{argmin}_{m\in\mathcal{M}} \Big\{ \ell\big(s^\star, \widehat{s}_m(D_n)\big) \Big\} \ .$$

Since $m^\star$ depends on the unknown distribution P, one can only hope to select $\widehat{m}(D_n)$ such that $\widehat{s}_{\widehat{m}}$ is almost as close to $s^\star$ as $\widehat{s}_{m^\star}$, which can be formalized as follows. An estimator selection procedure $\widehat{m}$ satisfies an *oracle inequality* with (leading) constant $C_n \geq 1$ and remainder term $R_n \geq 0$ when

$$\ell\big(s^\star, \widehat{s}_{\widehat{m}(D_n)}(D_n)\big) \leq C_n \inf_{m\in\mathcal{M}} \Big\{ \ell\big(s^\star, \widehat{s}_m(D_n)\big) \Big\} + R_n \tag{2.2}$$

holds either in expectation or with large probability (that is, a probability larger than $1 - C'/n^2$, for some constant $C' > 0$). Note that the oracle is often defined as

$$\operatorname*{argmin}_{m\in\mathcal{M}} \Big\{ \mathbb{E}\Big[\ell\big(s^\star, \widehat{s}_m(D_n)\big) \Big] \Big\} ,$$

leading to a weaker form of oracle inequality

$$\mathbb{E}\Big[\ell\big(s^\star, \widehat{s}_{\widehat{m}(D_n)}(D_n)\big)\Big] \leq C_n \inf_{m\in\mathcal{M}}\Big\{\mathbb{E}\Big[\ell\big(s^\star, \widehat{s}_m(D_n)\big)\Big]\Big\} + R_n \ .$$

Let us finally mention that in the statistical literature, the term oracle inequality sometimes refer to slightly different theoretical guarantees on statistical procedures [1].

In the asymptotic framework, if Eq. (2.2) holds on a large probability event with C_n tending to 1 when n tends to infinity and $R_n \ll \ell(s^\star, \widehat{s}_{m^\star}(D_n))$, then

$$\frac{\ell\big(s^\star, \widehat{s}_{\widehat{m}(D_n)}(D_n)\big)}{\inf_{m\in\mathcal{M}_n}\Big\{\ell\big(s^\star, \widehat{s}_m(D_n)\big)\Big\}} \xrightarrow[n\to\infty]{a.s.} 1$$

and the estimator selection procedure $\widehat{m}$ is called *efficient* (or asymptotically optimal).

In the non-asymptotic framework that we consider in the following, formally defining the optimality of an estimator selection procedure $\widehat{m}$ is more difficult. For instance, in Eq. (2.2), one can always decrease C_n by accordingly increasing R_n, and conversely. Assuming the remainder term R_n is indeed negligible in front of $\inf_{m\in\mathcal{M}}\{\ell(s^\star, \widehat{s}_m(D_n))\}$, an oracle inequality such as Eq. (2.2) is said to be *optimal* when the leading constant C_n is as small as possible, for a given family $\mathcal{M}$ and a set of possible distributions P. Because of the constraint $C_n \geq 1$, an oracle inequality (2.2) is optimal at first order if $C_n = 1 + \delta_n$ only depends on n and if $\delta_n \to 0$ as $n \to \infty$ (assuming again that R_n truly is a remainder term).

Building a procedure $\widehat{m}$ that satisfies an oracle inequality (2.2) is not only useful for the practical situation where a collection of estimators is given and one must choose among them. It can also be used for building *minimax adaptive* estimators, provided the family $(\widehat{s}_m)_{m\in\mathcal{M}}$ is well-chosen, see for instance [**BM97**, **BBM99**].

Let us conclude this section by describing the main challenge of estimator selection, that is, to avoid both overfitting and underfitting. For simplicity, let us consider the model selection problem with a family of models $(S_m)_{m\in\mathcal{M}}$. On the one hand, when S_m is "too small", any $t \in S_m$ is a poor approximation to $s^\star$, so that

$$\ell\big(s^\star, \widehat{s}_m(D_n)\big) \geq \inf_{t\in S_m}\big\{\ell(s^\star, t)\big\} := \ell(s^\star, S_m)$$

is large for most $s^\star \in \mathbb{S}$. The lower bound $\ell(s^\star, S_m)$ is called the *approximation error* or *bias* of model S_m. Thinking of nested models, $\ell(s^\star, S_m)$ is a nonincreasing function

1. Remark that in the definition (2.2) of an oracle inequality, we consider the relative loss $\ell(s^\star, \cdot)$ and *not* the loss $\mathcal{L}(\cdot)$. This choice is important: a guarantee of the form "the loss is smaller than twice the loss of the oracle" is often meaningless because the minimal value of the loss $\mathcal{L}(s^\star)$ is positive. So, even if the oracle is consistent, that is, if its loss converges to $\mathcal{L}(s^\star)$ as n goes to infinity, such a guarantee does not imply the selected estimator is consistent. It only implies that the loss of the selected estimator is asymptotically smaller than $2\mathcal{L}(s^\star) > \mathcal{L}(s^\star)$. This is the reason why we focus on the relative loss $\ell(s^\star, \cdot)$ instead of the loss, here and in the following.

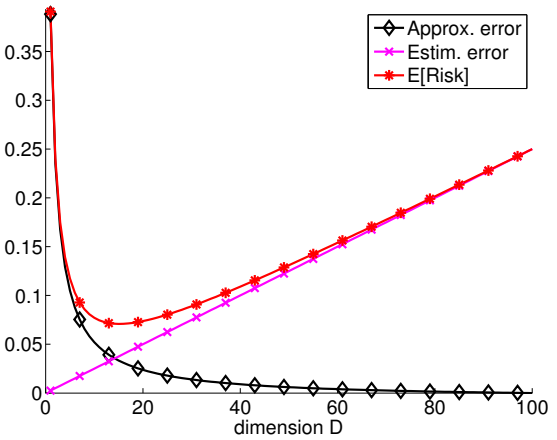


FIGURE 1. Illustration of the decomposition (2.3) of $\mathbb{E}[\ell(s^\star, \widehat{s}_m(D_n))]$ (red stars) into the sum of the approximation error (black diamonds) and the estimation error (blue crosses), plotted as a function of the dimension D_m of the models, in a fixed-design regression setting with $n = 100$ data points and a constant noise level $\sigma^2 = 1/4$; see also Section 4.2 and **[22]**.

of S_m. On the other hand, when S_m is "too large", $\widehat{s}_m(D_n)$ is likely to overfit: this results from the *estimation error*. Think for instance of S_m as a parametric model with more than n parameters, or the set of all continuous functions on $[0, 1]$ in the regression framework.

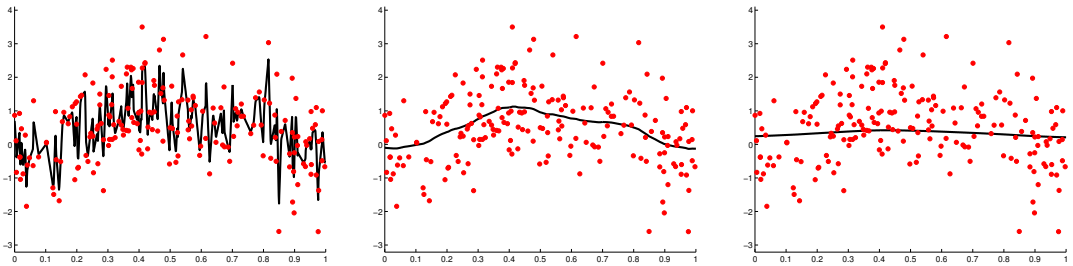
If S_m is a vector space of dimension D_m, it can be proved in several classical frameworks that

$$\begin{aligned} \mathbb{E}\Big[\ell\big(s^\star, \widehat{s}_m(D_n)\big)\Big] &= \text{Approximation error} + \text{Estimation error} \\ &\approx \ell(s^\star, S_m) + \alpha_n D_m \ , \end{aligned} \tag{2.3}$$

where $\alpha_n > 0$ does not depend on m. For instance, $\alpha_n = 1/(2n)$ in density estimation using the log-likelihood contrast, and $\alpha_n = \sigma^2/n$ in regression using the least-squares contrast and assuming that $\mathrm{var}(Y \mid X) = \sigma^2$ does not depend on X. See also Eq. (4.5) in Section 4.3 and Eq. (4.28) in Section 4.6.

According to Eq. (2.3), a good model choice must reach the best trade-off between the *approximation error* $\ell(s^\star, S_m)$ and the *estimation error* $\alpha_n D_m$, which is often called the *bias-variance trade-off*—the term $\alpha_n D_m$ is often called "variance"; see also Figure 1.

A similar phenomenon holds in the general case, in particular hyperparameter tuning; think for instance of the problem of choosing k for k-nearest neighbors, or of the bandwidth selection problem for kernel density estimation. Even if the risk cannot always be decomposed into approximation and estimation errors, one always wants to avoid overfitting—that is, following the data too closely, which prevents from generalizing because data are noisy—and underfitting—that is, considering "too simple"



(a) overfitting (b) intermediate situation (c) underfitting

FIGURE 2. Illustration of the overfitting and underfitting phenomena in regression with kernel ridge estimators. Data (red points) $(X_i, Y_i)_{1 \leq i \leq n}$ are generated by $Y_i = \sin(\pi X_i) + \varepsilon_i$, $i = 1, \dots, n = 200$, where the X_i are independent with uniform distribution over $[0, 1]$ and the ε_i are independent (and independent from the X_i) with a standard normal distribution. The three estimators of $s^\star : x \mapsto \sin(\pi x)$ considered (black curves) are kernel ridge density estimators with the Laplace kernel and a regularization parameter λ equal to (a) 10^{-5}, (b) 10^{-2} and (c) $1/2$.

estimators, that cannot describe well the underlying signal. These extreme situations, as well as an intermediate one which realizes the desired trade-off, can be visualized on Figure 2.

2.2. General approach to estimator selection

The most classical estimator selection procedures—and in particular those that we consider in the following—are defined by

$$\widehat{m} = \widehat{m}_{\mathcal{C}} \in \operatorname*{argmin}_{m \in \mathcal{M}} \big\{ \mathcal{C}(m) \big\} \tag{2.4}$$

for some (data-driven) criterion $\mathcal{C} : \mathcal{M} \to \mathbb{R}$ such as cross-validation (see Chapter 3) or a penalized empirical criterion (see Chapter 4). This section describes how such procedures can be analyzed theoretically, which provides some principles for building estimator selection procedures that work well in practice. We refer to [**BBM99**, **BA02**, **Mas07**, **HTF09**, **BGH10**, **Gir14**] for a bibliography on model and estimator selection.

2.2.1. Estimator selection as a particular case of a general problem. Given a procedure defined by Eq. (2.4), as described in Section 2.1.5, the goal is to prove an oracle inequality such as Eq. (2.2), that is, to upper-bound the excess loss of the final estimator

$$\ell\big(s^\star, \widehat{s}_{\widehat{m}_{\mathcal{C}}(D_n)}(D_n)\big) = \mathcal{R}(\widehat{m}_{\mathcal{C}}) \quad \text{where} \quad \forall m \in \mathcal{M}, \quad \mathcal{R}(m) := \ell(s^\star, \widehat{s}_m) \ .$$

Introducing the notation $\mathcal{R}(\cdot)$ allows to cast estimator selection as an instance of the following general problem—which is central to statistics, learning and optimization—,

hence allowing some comparisons:

Try to minimize $\mathcal{R}(m)$ over $m \in \mathcal{M}$ by minimizing $\mathcal{C}(m)$ over $m \in \mathcal{M}$ instead. (2.5)

In statistics, minimum contrast estimators also are an instance of problem (2.5): if $\mathcal{M} = S \subset \mathbb{S}$ is some model and if for every $t \in S$, we define $\mathcal{C}(t) = P_n\gamma(t)$ the empirical risk, then $\widehat{m}_{\mathcal{C}} = \widehat{s}_S$ is a minimum contrast estimator over S and the goal is to minimize $\mathcal{R}(t) = \ell(s^\star, t)$. Another example of problem (2.5) is the use of relaxations in optimization: $\mathcal{R}(m)$ can be computed for every $m \in \mathcal{M}$ but minimizing it exactly over $\mathcal{M}$ is not tractable, so instead the (easier) optimization problem, "minimize $\mathcal{C}(m)$ over $m \in \mathcal{M}$" is solved, for instance by considering a convex relaxation $\mathcal{C}$ of $\mathcal{R}$. Classifiers based on convex losses combine these two problems: The 0–1 contrast is replaced by a convex surrogate γ, and the loss $P\gamma(t)$ is replaced by its empirical counterpart $P_n\gamma(t)$.

A classical way to analyze (2.5)—in particular minimum contrast estimators and estimator selection procedures—is summarized by the following lemma.

Lemma 2.1. *Let $\mathcal{R}, \mathcal{C}, A, B$ be some mappings $\mathcal{M} \to \mathbb{R}$, possibly data-dependent. Then, on the event Ω on which*

$$\forall m, m' \in \mathcal{M}, \quad \big(\mathcal{C}(m) - \mathcal{R}(m)\big) - \big(\mathcal{C}(m') - \mathcal{R}(m')\big) \leq A(m) + B(m') \ , \tag{2.6}$$

we have

$$\forall \widehat{m}_{\mathcal{C}} \in \operatorname*{argmin}_{m \in \mathcal{M}} \big\{\mathcal{C}(m)\big\}, \quad \mathcal{R}(\widehat{m}_{\mathcal{C}}) - B(\widehat{m}_{\mathcal{C}}) \leq \inf_{m \in \mathcal{M}} \big\{\mathcal{R}(m) + A(m)\big\} \ . \tag{2.7}$$

In particular, Eq. (2.7) holds true on the event $\Omega' \subset \Omega$ on which

$$\forall m \in \mathcal{M}, \quad -B(m) \leq \mathcal{C}(m) - \mathcal{R}(m) \leq A(m) \ . \tag{2.8}$$

We remark that quantities of the form $\mathcal{C}(m) - \mathcal{C}(m')$ appear in relative bounds [**Cat07**, Section 1.4] which can be used as a tool for model selection [**Aud04**].

Proof of Lemma 2.1. Assume Eq. (2.6) holds true. By definition of $\widehat{m}_{\mathcal{C}}$ in Eq. (2.7), for every $m \in \mathcal{M}$,

$$\mathcal{C}(\widehat{m}_{\mathcal{C}}) \leq \mathcal{C}(m) \ ,$$

so that

$$\mathcal{R}(\widehat{m}_{\mathcal{C}}) - \mathcal{R}(m) \leq \mathcal{R}(\widehat{m}_{\mathcal{C}}) - \mathcal{R}(m) + \underbrace{\mathcal{C}(m) - \mathcal{C}(\widehat{m}_{\mathcal{C}})}_{\geq 0} \leq A(m) + B(\widehat{m}_{\mathcal{C}})$$

by Eq. (2.6) with $m' = \widehat{m}_{\mathcal{C}}$. Reordering the terms gives Eq. (2.7). The last result is straightforward: Eq. (2.8) implies Eq. (2.6). □

For estimator selection, Lemma 2.1 provides simple sufficient conditions for an oracle inequality (2.2) to hold. If, for all $m, m' \in \mathcal{M}$, the difference $\mathcal{C}(m) - \mathcal{C}(m')$ estimates

"well" $\mathcal{R}(m) - \mathcal{R}(m')$, that is, Eq. (2.6) holds true with

$$\sup_{m \in \mathcal{M}} \left\{ \frac{\max\{A(m), B(m)\}}{\mathcal{R}(m)} \right\} \leq \delta < 1 \ , \tag{2.9}$$

then Eq. (2.7) can be rewritten as the following oracle inequality:

$$\mathcal{R}(\widehat{m}_{\mathcal{C}}) \leq \frac{1+\delta}{1-\delta} \inf_{m \in \mathcal{M}} \{\mathcal{R}(m)\} \ ,$$

which is optimal at first order if $\delta = \delta_n = \mathrm{o}(1)$ as n tends to infinity.

Since Eq. (2.8) is a sufficient condition to Eq. (2.6), Lemma 2.1 suggests a classical strategy for building estimator selection procedures, that we comment on in Section 2.2.2: use a criterion $\mathcal{C}(m)$ that unbiasedly estimates $\mathbb{E}[\mathcal{R}(m)]$ for every $m \in \mathcal{M}$.

Let us now consider another situation, assuming that $\mathcal{C}$ is a uniform upper bound on $\mathcal{R}$ over $\mathcal{M}$, that is, Eq. (2.8) holds true with $B(m) = 0$. Then, Eq. (2.7) can be rewritten as

$$\mathcal{R}(\widehat{m}_{\mathcal{C}}) \leq \inf_{m \in \mathcal{M}} \{\mathcal{R}(m) + A(m)\} \tag{2.10}$$

which is an oracle inequality provided $A(m)$ is not much larger than $\mathcal{R}(m)$. The corresponding strategy is used in particular for "large" families of estimators, as described in Section 2.2.3.

In particular, in the framework of Section 2.1.2, penalization methods correspond to taking

$$\mathcal{C}(m) = P_n \gamma(\widehat{s}_m) + \mathrm{pen}(m) \tag{2.11}$$

for some function $\mathrm{pen} : \mathcal{M} \to \mathbb{R}$, called the penalty. Then, if $\mathcal{R}(m) = \ell(s^\star, \widehat{s}_m)$,

$$\mathcal{C}(m) - \mathcal{R}(m) = \mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m) + P\gamma(s^\star)$$
$$\text{where} \qquad \mathrm{pen}_{\mathrm{id}}(m) := (P - P_n)\gamma(\widehat{s}_m)$$

is the "ideal penalty", since using it would lead to minimizing $\mathcal{R}(m)$ exactly. So, for penalization methods, Lemma 2.1 can be rewritten with $\mathrm{pen} - \mathrm{pen}_{\mathrm{id}}$ instead of $\mathcal{C} - \mathcal{R}$ in the sufficient conditions (2.6) or (2.8). This suggests to use a penalty that either unbiasedly estimates $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$ or is a uniform upper bound on $\mathrm{pen}_{\mathrm{id}}$.

Let us conclude this subsection by showing how the classical analysis of minimum contrast estimators follows the lines of Lemma 2.1. Recall that $\mathcal{M} = S \subset \mathbb{S}$ is a model and for every $t \in S$, $\mathcal{C}(t) = P_n\gamma(t)$ and $\mathcal{R}(t) = \ell(s^\star, t)$, so that $\widehat{m}_{\mathcal{C}} = \widehat{s}_S$ is a minimum contrast estimator over S. For every $c \in \mathbb{R}$, Eq. (2.6) always holds true with

$$A(m) = B(m') = \sup_{m'' \in \mathcal{M}} \left| \mathcal{R}(m'') - \mathcal{C}(m'') + c \right| \ .$$

Lemma 2.1 shows that

$$\mathcal{R}(\widehat{m}_{\mathcal{C}}) \leq \inf_{m \in \mathcal{M}} \{\mathcal{R}(m)\} + 2 \sup_{m'' \in \mathcal{M}} \left| \mathcal{R}(m'') - \mathcal{C}(m'') + c \right|$$

which can be rewritten as

$$\ell(s^\star, \widehat{s}_S) \leq \inf_{t\in S}\big\{\ell(s^\star,t)\big\} + 2\sup_{t\in S}\big|(P-P_n)\gamma(t)\big| = \ell(s^\star,S) + 2\sup_{t\in S}\big|(P-P_n)\gamma(t)\big| \quad (2.12)$$

by taking $c = P\gamma(s^\star)$. Eq. (2.12) corresponds to the first step towards global risk bounds for empirical risk minimization in classification, see for instance [**BBL05**, Section 3].

Taking A and B constant in Eq. (2.6) can be far too pessimistic. The so-called localization approach [**BBL05**, Section 5] improves the above analysis on this point, by taking into account that $\widehat{s}_S$ can be localized in a region of S where the deviations of $(P-P_n)\gamma(t)$ can be controlled more tightly than with a uniform bound. For instance, [**Mas07**, Section 8.3] shows that under the "margin condition" (see Section 6.4), some ratio-type inequality such as

$$\sup_{t\in S}\left\{\frac{(P-P_n)\big(\gamma(t)-\gamma(s_S^\star)\big)}{\ell(s^\star,t)+x^2}\right\} \leq \varepsilon \quad (2.13)$$

holds with a large probability, where

$$s_S^\star \in \operatorname*{argmin}_{t\in S}\big\{\ell(s^\star,t)\big\}$$

and $\varepsilon, x > 0$ are fixed. Coming back to the notation of Lemma 2.1, Eq. (2.13) can be rewritten as

$$\sup_{m\in\mathcal{M}}\left\{\frac{\big(\mathcal{R}(m)-\mathcal{C}(m)\big)-\big(\mathcal{R}(m^\star)-\mathcal{C}(m^\star)\big)}{\mathcal{R}(m)+x^2}\right\} \leq \varepsilon \quad \text{where} \quad m^\star \in \operatorname*{argmin}_{m\in\mathcal{M}}\mathcal{R}(m)\ ,$$

which implies that Eq. (2.6) holds true with

$$B(m) = \varepsilon\big(\mathcal{R}(m)+x^2\big)\,, \qquad A(m^\star) = 0 \qquad \text{and} \quad \forall m' \neq m^\star,\ A(m') = +\infty\,.$$

Lemma 2.1 shows that this implies

$$(1-\varepsilon)\mathcal{R}(\widehat{m}_{\mathcal{C}}) \leq \mathcal{R}(m^\star) + \varepsilon x^2$$

hence, assuming $\varepsilon \in (0,1)$,

$$\ell(s^\star, \widehat{s}_S) \leq \frac{1}{1-\varepsilon}\ell(s^\star,S) + \frac{\varepsilon}{1-\varepsilon}x^2\ .$$

Interestingly, this argument is quite similar to proving that Eq. (2.9) holds true for an estimator selection procedure, in order to derive an oracle inequality.

2.2.2. First-order optimality and the unbiased risk estimation principle. As explained in the previous subsection, a classical and widely used strategy for estimator selection is the "unbiased risk estimation principle": minimize a criterion $\mathcal{C}$ such that $\mathcal{C}(m)$ estimates unbiasedly the risk $\mathbb{E}[\mathcal{R}(m)] = \mathbb{E}[\ell(s^\star, \widehat{s}_m)]$ for every $m \in \mathcal{M}$. Equivalently, for penalized criteria defined by Eq. (2.11), this principle suggests to use a penalty that unbiasedly estimates the expectation of the ideal penalty $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)] = \mathbb{E}[(P-P_n)\gamma(\widehat{s}_m)]$.

This principle, also known as Mallows' or Akaike's heuristics, leads to many estimator selection procedures such as cross-validation (see Chapter 3), Akaike's Information Criterion [**Aka73**, AIC] and Mallows' C_p [**Mal73**] (see Chapter 4). More examples can be found in Section 3 of **[7]**.

Given Lemma 2.1 and assuming $\mathbb{E}[\mathcal{C}(m)] = \mathbb{E}[\mathcal{R}(m)]$ for every $m \in \mathcal{M}$, for proving an oracle inequality, it is sufficient to prove Eq. (2.8) holds with a large probability. In other words, what remains is to prove a concentration inequality for $\mathcal{C}(m) - \mathcal{R}(m)$ around its expectation (zero), that holds simultaneously for all $m \in \mathcal{M}$ on a large probability event, with small enough deviation bounds—for instance, satisfying Eq. (2.9). In several settings, under suitable assumptions on the data, such a concentration inequality holds true for every single $m \in \mathcal{M}$. Then, a union bound can make this concentration result uniform over $m \in \mathcal{M}$ provided there are "not too many estimators", for instance, if $\mathcal{M}$ is "polynomial", which can be defined as $\mathrm{Card}(\mathcal{M}) \leq Ln^{\alpha}$ for some numerical constants $L, \alpha > 0$. A more precise definition of a "polynomial" family $\mathcal{M}$ in the context of model selection can be found in [**BM07**]. Note that $\mathcal{M}$ can be "small enough" while being infinite. For instance, **[18]** shows that the family $(\widehat{s}_m)_{m \in \mathcal{M}}$ of kernel ridge estimators for some fixed kernel but a varying regularization parameter $m \in \mathcal{M} = \mathbb{R}$ is small in this sense, although it is infinite.

Given such a concentration result with small enough deviation bounds, say Eq. (2.9) with $\delta = \delta_n = \mathrm{o}(1)$ as n goes to infinity, Lemma 2.1 implies a first-order optimal oracle inequality for $\widehat{m}_{\mathcal{C}}$. Therefore, when analyzing an estimator selection method at first order, the key problem is to compute—or to approximate—the expectation of the criterion, $\mathbb{E}[\mathcal{C}(m)]$, and to compare it with the expectation of the quantity we want to minimize, $\mathbb{E}[\mathcal{R}(m)]$. This point appears clearly in the following, in particular in the analysis of cross-validation and resampling methods in Chapter 3 and of penalization procedures in Chapter 4, in particular in the proof presented in Section 4.5.

2.2.3. Large collections of estimators. When $\mathcal{M}$ is "large"—for instance, "exponential", as defined in [**BM07**]—, the unbiased risk estimation principle of Section 2.2.2 breaks down. It is usually replaced by using a *uniform* upper bound on $\mathcal{R}(m)$ as criterion $\mathcal{C}(m)$. Equivalently, for penalization procedures defined by Eq. (2.11), the penalty is chosen in order to satisfy $\mathrm{pen}(m) \geq \mathrm{pen}_{\mathrm{id}}(m)$ *simultaneously* for all $m \in \mathcal{M}$. Then, from Lemma 2.1, we get an oracle inequality of the form of Eq. (2.10). Clearly, the tighter is the upper bound on $\mathcal{R}(m)$, the better are the theoretical guarantees. A principled way of building such an upper bound is explained in [**BM01**] for model selection, through the choice of weights $(L_m)_{m \in \mathcal{M}}$ that can reflect prior knowledge on the estimators $(\widehat{s}_m)_{m \in \mathcal{M}}$. This approach has successful applications, in particular for change-point detection, see Chapter 5.

2.2.4. Comparing estimator selection procedures. Let us make a few comments about a key problem mentioned in introduction: How to compare estimator selection procedures, that is, given two criteria $\mathcal{C}_1$ and $\mathcal{C}_2$, how to compare the performances of $\widehat{m}_{\mathcal{C}_1}$ and $\widehat{m}_{\mathcal{C}_2}$?

Ideally, for proving that $\mathcal{C}_1$ is better than $\mathcal{C}_2$ in some setting, we would like to prove that

$$\ell\big(s^\star, \widehat{s}_{\widehat{m}_{\mathcal{C}_1}}\big) < (1-\varepsilon_n)\,\ell\big(s^\star, \widehat{s}_{\widehat{m}_{\mathcal{C}_2}}\big) \tag{2.14}$$

with a large probability, for some $\varepsilon_n \geq 0$; see also [**Yan07**, Definition 1] for a slightly more precise formulation in the asymptotic framework.

The classical way to analyze the performance of an estimator selection procedure $\widehat{m}_{\mathcal{C}}$ is to prove an oracle inequality such as Eq. (2.2), that is, to *upper-bound* (with a large probability or in expectation)

$$\ell\big(s^\star, \widehat{s}_{\widehat{m}_{\mathcal{C}}}\big) - \inf_{m\in\mathcal{M}_n}\big\{\ell(s^\star, \widehat{s}_m)\big\} \quad \text{or} \quad \mathfrak{R}_n(\mathcal{C}) := \frac{\ell\big(s^\star, \widehat{s}_{\widehat{m}_{\mathcal{C}}}\big)}{\inf_{m\in\mathcal{M}_n}\big\{\ell(s^\star, \widehat{s}_m)\big\}} \ .$$

Alternatively, asymptotic results show that when n tends to infinity, $\mathfrak{R}_n(\mathcal{C}) \to 1$ (asymptotic optimality of $\mathcal{C}$) or that $\mathfrak{R}_n(\mathcal{C}_1) \sim \mathfrak{R}_n(\mathcal{C}_2)$ (asymptotic equivalence of $\mathcal{C}_1$ and $\mathcal{C}_2$). Nevertheless, comparing the bounds we can prove for $\mathcal{C}_1$ and for $\mathcal{C}_2$ might not reflect the actual ordering between $\mathcal{C}_1$ and $\mathcal{C}_2$. Given an oracle inequality for $\mathcal{C}_1$, in order to prove a result such as Eq. (2.14), we need a *lower bound* on $\mathfrak{R}_n(\mathcal{C}_2)$.

A usual approach is to prove such lower bounds in the worst case, that is, minimax lower bounds, as for instance in **[8]**. Such lower bounds are not necessarily informative for practitioners if the worse-case scenario is not realistic, or if the worst-case scenario is not explicitly described in the minimax analysis.

Therefore, we believe there is a need for theoretical negative results that hold under mild assumptions, or at least in some explicit and realistic particular cases. Proving such non-asymptotic lower bounds is usually difficult, because matching the known upper bounds requires a very precise analysis.

Yet, this is sometimes possible. As suggested by Section 2.2.2, a comparison of $\mathcal{C}_1$ and $\mathcal{C}_2$ at first order relies primarily on a comparison of their expectations. The simplest case is when some decomposition such as Eq. (2.3) is available for $\mathcal{R}(m)$ and in addition

$$\forall i \in \{1,2\}, \quad \mathbb{E}\big[\mathcal{C}_i(m)\big] = \ell(s^\star, S_m) + \kappa_i \alpha_n D_m$$

for some constants $\kappa_1 \neq \kappa_2$. Then, up to some concentration inequalities, a comparison of $\mathcal{C}_1$ and $\mathcal{C}_2$ is possible under relatively mild assumptions, as done for instance for some cross-validation methods in **[16]** (see also Chapter 3) and for some penalization procedures in Chapter 4. When the shapes of $\mathbb{E}[\mathcal{R}(m)]$, $\mathbb{E}[\mathcal{C}_1(m)]$ and $\mathbb{E}[\mathcal{C}_2(m)]$ are less similar, such comparisons can still be possible, as done for instance in **[17]** for some penalization procedures in heteroscedastic regression, by considering a more specific setting and designing a proof specially for this setting.

2.2.5. Beyond first order: taking into account the variance. The major limitation of the approach of Sections 2.2.2–2.2.4 is that it can only compare $\mathcal{C}_1$ to $\mathcal{C}_2$ at first order, that is, according to

$$\lim_{n\to\infty} \frac{\mathfrak{R}_n(\mathcal{C}_1)}{\mathfrak{R}_n(\mathcal{C}_2)},$$

which usually depends only on the bias of $\mathcal{C}_i(m)$ $(i = 1, 2)$ as an estimator of $\mathbb{E}[\mathcal{R}(m)]$. For instance, leave-$p$-out and hold-out with a training set of size $n - p$ cannot be distinguished at first order, as explained in Section 6 of [7], while leave-p-out performs much better in practice, certainly because its "variance" is much smaller, see Chapter 3.

So, we must go beyond the first order of $\mathfrak{R}_n(\mathcal{C})$ and take into account the variance of $\mathcal{C}(m)$. Nevertheless, proving a lower bound on $\mathfrak{R}_n(\mathcal{C})$ is already challenging at first order, and second-order terms are clearly not precise in most non-asymptotic upper bounds. As a first step, we propose a heuristics showing the variances of some quantities—depending on $(\mathcal{C}_i)_{i=1,2}$ and on $\mathcal{M}$—can be used as a proxy to a proper comparison of the second-order terms of $\mathfrak{R}_n(\mathcal{C}_1)$ and $\mathfrak{R}_n(\mathcal{C}_2)$. Since we focus on second-order terms, from now on, we assume $\mathcal{C}_1$ and $\mathcal{C}_2$ have the same bias, that is,

$$\forall m \in \mathcal{M}, \quad \mathbb{E}\big[\mathcal{C}_1(m)\big] = \mathbb{E}\big[\mathcal{C}_2(m)\big] \ . \tag{\textbf{SameBias}}$$

If we were only comparing $\mathcal{C}_1$ and $\mathcal{C}_2$ as estimators of $\mathbb{E}[\mathcal{R}(m)]$ for every single $m \in \mathcal{M}$, we could naturally compare them through their mean squared errors. Under assumption (**SameBias**), this would mean to compare their variances. This is not sufficient to solve our problem, because risk estimation and estimator selection are different tasks [**BS92**]. For instance, $\widehat{m}_{\mathcal{C}}$ defined by Eq. (2.4) is unchanged when $\mathcal{C}(m)$ is translated by any random quantity, but such a translation does change $\mathrm{var}\big(\mathcal{C}(m)\big)$ and can make it as large as desired. For estimator selection—and more generally for problem (2.5), what really matters is that

$$\mathrm{sign}\big(\mathcal{C}(m_1) - \mathcal{C}(m_2)\big) = \mathrm{sign}\big(\mathcal{R}(m_1) - \mathcal{R}(m_2)\big) \tag{2.15}$$

as often as possible for every $m_1, m_2 \in \mathcal{M}$, and that most mistakes in the ranking of estimators occur when $\mathcal{R}(m_1) - \mathcal{R}(m_2)$ is small, so that $\mathcal{R}(\widehat{m}_{\mathcal{C}})$ cannot be much larger than $\inf_{m\in\mathcal{M}}\big\{\mathcal{R}(m)\big\}$.

The heuristics we propose goes as follows. Assume for simplicity that

$$m^\star = \operatorname*{argmin}_{m\in\mathcal{M}}\Big\{\mathbb{E}\big[\mathcal{R}(m)\big]\Big\}$$

is uniquely defined. For any $\mathcal{C}$, the more concentrated "around $m^\star$" is the distribution of $\widehat{m}_{\mathcal{C}}$, the better is the performance of $\widehat{s}_{\widehat{m}_{\mathcal{C}}}$. Let us now simplify this fact into "$\mathcal{C}$ is better if $\mathbb{P}(m = \widehat{m}_{\mathcal{C}})$ is smaller for all $m \neq m^\star$". Our idea is to find a proxy for $\mathbb{P}(m = \widehat{m}_{\mathcal{C}})$, that is, a quantity that should behave similarly as a function of $\mathcal{C}$ and its "variance" properties. For all $m, m' \in \mathcal{M}$, let us define $\Delta_{\mathcal{C}}(m, m') := \mathcal{C}(m) - \mathcal{C}(m')$ and $\overline{\Phi}(t) = \mathbb{P}(\xi > t)$ for all $t \in \mathbb{R}$, and let ξ be a standard Gaussian random variable. Then,

for every $m \in \mathcal{M}$,

$$\begin{aligned}
&\mathbb{P}(\widehat{m}_{\mathcal{C}} = m) = \mathbb{P}\big(\forall m' \neq m\,,\, \Delta_{\mathcal{C}}(m,m') < 0\big) \\
&\asymp \min_{m' \neq m} \mathbb{P}\big(\Delta_{\mathcal{C}}(m,m') < 0\big) \qquad (2.16) \\
&\approx \min_{m' \neq m} \mathbb{P}\Big(\mathbb{E}\big[\Delta_{\mathcal{C}}(m,m')\big] + \xi\sqrt{\mathrm{var}\big(\Delta_{\mathcal{C}}(m,m')\big)} < 0\Big) \qquad (2.17) \\
&= \overline{\Phi}\big(\mathrm{SNR}_{\mathcal{C}}(m)\big) \quad \text{where} \quad \mathrm{SNR}_{\mathcal{C}}(m) := \max_{m' \neq m} \frac{\mathbb{E}\big[\Delta_{\mathcal{C}}(m,m')\big]}{\sqrt{\mathrm{var}\big(\Delta_{\mathcal{C}}(m,m')\big)}}
\end{aligned}$$

is the "signal-to-noise ratio" of increments of $\mathcal{C}$. So, if $\mathrm{SNR}_{\mathcal{C}_1}(m) > \mathrm{SNR}_{\mathcal{C}_2}(m)$ for all $m \neq m^\star$, $\mathcal{C}_1$ should be better than $\mathcal{C}_2$. Under assumption (**SameBias**), this leads to the following heuristics:

$$\left.\begin{array}{l} \text{If} \quad \forall m \neq m', \quad \mathrm{var}\big(\mathcal{C}_1(m) - \mathcal{C}_1(m')\big) < \mathrm{var}\big(\mathcal{C}_2(m) - \mathcal{C}_2(m')\big)\,, \\ \text{then,} \quad \mathcal{C}_1 \text{ is better than } \mathcal{C}_2\ . \end{array}\right\} \qquad (2.18)$$

Let us make some remarks.

- Approximation (2.16) is the strongest one. Clearly, inequality $\leq$ holds true. The equality case is for a very particular dependence setting, when the events
$$\Big(\big\{\Delta_{\mathcal{C}}(m,m') < 0\big\}\Big)_{m' \in \mathcal{M}}$$
are nested. In general, the left-hand side is significantly smaller than the right-hand side; we claim that they vary similarly as a function of $\mathcal{C}$.
- The Gaussian approximation (2.17) for $\Delta_{\mathcal{C}}(m,m')$ does not hold exactly, but it seems reasonable to make it, at first order at least.
- Approximations (2.16) and (2.17) are tested numerically in **[14]** for cross-validation methods in least-squares density estimation (see Chapter 3), showing the above heuristics provides reasonably good comparisons.

In the heuristics (2.18), all (m,m') do not matter equally for explaining a quantitative difference in the performance of $\mathcal{C}$. First, we can fix $m' = m^\star$, since intuitively the strongest candidate against any $m \neq m^\star$ is $m^\star$. Second, if m and $m^\star$ are very close, that is, $\mathcal{R}(m)/\mathcal{R}(m^\star)$ is smaller than the minimal order of magnitude we can expect for $\mathfrak{R}_n(\mathcal{C})$ with a data-driven $\mathcal{C}$, taking m instead of $m^\star$ does not decrease the performance significantly. Third, if $\overline{\Phi}\big(\mathrm{SNR}_{\mathcal{C}}(m)\big)$ is very small, changing it even by an order of magnitude cannot significantly affect the performance of $\widehat{m}_{\mathcal{C}}$; hence, all m such that, say, $\mathrm{SNR}_{\mathcal{C}}(m) \gg \big(\log(n)\big)^\alpha$ for all $\alpha > 0$ can also be discarded. Overall, pairs (m,m') that really matter in (2.18) are pairs $(m,m^\star)$ that are at a "moderate distance" in terms of $\mathbb{E}[\mathcal{R}(m) - \mathcal{R}(m^\star)]$.

CHAPTER 3

Cross-validation and resampling

Cross-validation (CV) methods are popular in statistics and in machine learning, for estimating the loss of a given estimator and for estimator selection. In short, the CV principle is to split—once or several times—the data into a training sample, that is used for training estimators, and a validation sample, that is used for estimating the loss of the trained estimators by measuring how they perform on "new" data points. CV can thus be seen as an application of subsampling, which is itself part of the general idea of *resampling*, first introduced with the bootstrap [**Efr79**].

Few non-asymptotic theoretical results are available on CV or resampling methods. In particular, when several CV methods are compared, most theoretical results are not precise enough to explain the relative behaviors which can be observed in practice. This chapter summarizes several works that primarily aim at narrowing this gap between theoretical knowledge and empirical observations on CV and resampling methods for the estimator selection problem: **[16]**, **[3]** and **[14]**. More references and a more complete picture of the current knowledge about CV for model/estimator selection can be found in the survey paper **[7]**. Finally, in Section 3.4, we briefly mention other works on resampling and CV methods: **[2]**, **[10]**, **[4]** and **[5]**.

3.1. Definitions

First, let us recall the definitions of the most classical CV estimators of the risk of some estimator $\widehat{s}$, that can be used as a criterion $\mathcal{C}$ for defining an estimator selection procedure $\widehat{m}_{\mathcal{C}}$.

3.1.1. Hold-out. The simplest method is *hold-out* [**DW79**] or (simple) *validation*, which relies on a single split of data. Let $I^{(t)}$ be a non-empty proper subset of $\{1,\dots,n\}$, that is, such that both $I^{(t)}$ and its complement $I^{(v)} = (I^{(t)})^c = \{1,\dots,n\}\backslash I^{(t)}$ are non-empty. The *hold-out* estimator of the risk of $\widehat{s}(D_n)$, with *training set* $I^{(t)}$, is given by

$$\widehat{\mathcal{L}}^{\mathrm{HO}}\left(\widehat{s}; D_n; I^{(t)}\right) := P_n^{(v)}\gamma\left(\widehat{s}\left(D_n^{(t)}\right)\right) = \frac{1}{n_v}\sum_{i\in D_n^{(v)}}\gamma\left(\widehat{s}\left(D_n^{(t)}\right);\xi_i\right) \ , \tag{3.1}$$

where $D_n^{(t)} := (\xi_i)_{i\in I^{(t)}}$ is the *training sample*, of size $n_t = \mathrm{Card}(I^{(t)})$, $D_n^{(v)} := (\xi_i)_{i\in I^{(v)}}$ is the *validation sample*, of size $n_v = n - n_t$ and $P_n^{(v)}$ is its empirical distribution; $I^{(v)}$ is called the *validation set*.

3.1.2. General definition of cross-validation. A general description of the CV strategy has been given by [Gei75]. In brief, CV consists in averaging several hold-out estimators of the risk corresponding to different data splits. Let $B \geq 1$ be an integer and $I_1^{(t)}, \ldots, I_B^{(t)}$ a sequence of non-empty proper subsets of $\{1, \ldots, n\}$. The CV estimator of the risk of $\widehat{s}(D_n)$, with training sets $(I_j^{(t)})_{1 \leq j \leq B}$, is defined by

$$\begin{aligned} \widehat{\mathcal{L}}^{\mathrm{CV}}\Big(\widehat{s}; D_n; \big(I_j^{(t)}\big)_{1 \leq j \leq B}\Big) &:= \frac{1}{B} \sum_{j=1}^{B} \widehat{\mathcal{L}}^{\mathrm{HO}}\Big(\widehat{s}; D_n; I_j^{(t)}\Big) \\ &= \frac{1}{B} \sum_{j=1}^{B} P_n^{(v),j} \gamma\Big(\widehat{s}\big(D_n^{(t),j}\big)\Big) \ , \end{aligned} \tag{3.2}$$

$$\text{where} \qquad D_n^{(t),j} = (\xi_i)_{i \in I_j^{(t)}} \qquad \text{and} \qquad P_n^{(v),j} = \frac{1}{n - \mathrm{Card}\Big(I_j^{(t)}\Big)} \sum_{i \notin I_j^{(t)}} \delta_{\xi_i}$$

respectively denote the j-th training sample and the empirical distribution of the j-th validation sample. All classical CV estimators of the risk are of the form (3.2), usually with a fixed size n_t of the training set, that is,

$$\mathrm{Card}\Big(I_j^{(t)}\Big) \approx n_t$$

for every j. When $(I_j^{(t)})_{1 \leq j \leq B}$ is random, it is chosen independently from the data D_n. What remains is to choose n_t, B and the splitting scheme.

Given n_t, two main categories of splitting schemes can be distinguished: *exhaustive data splitting*, that is, considering all training sets of size n_t, and *partial data splitting*.

3.1.3. Exhaustive data splitting. The most classical exhaustive CV procedure is the leave-one-out procedure (LOO), which corresponds to the choice $n_t = n - 1$, as proposed by [Sto74, All74, Gei75]. Each data point is successively "left out" from the sample and used for validation. Formally, LOO is defined by Eq. (3.2) with $B = n$ and $I_j^{(t)} = \{j\}^c$ for $j = 1, \ldots, n$:

$$\widehat{\mathcal{L}}^{\mathrm{LOO}}(\widehat{s}; D_n) = \frac{1}{n} \sum_{j=1}^{n} \gamma\Big(\widehat{s}\big(D_n^{(-j)}\big); \xi_j\Big) \tag{3.3}$$

where $D_n^{(-j)} = (\xi_i)_{1 \leq i \leq n, i \neq j}$.

Its natural extension is the leave-p-out procedure (LPO) [Sha93] for some integer $p \in \{1, \ldots, n-1\}$, which is the exhaustive CV with $n_t = n - p$. Every possible subset of size p is successively "left out" from the sample and used for validation. Therefore, LPO is defined by (3.2) with

$$B = \binom{n}{p} \qquad \text{and} \qquad \Big(I_j^{(t)}\Big)_{1 \leq j \leq B} = \mathfrak{P}_{n-p}\big(\{1, \ldots, n\}\big)$$

is the set of all parts of $\{1,\dots,n\}$ of size $n-p$.

3.1.4. Partial data splitting. Considering $\binom{n}{p}$ training sets can be computationally intractable, even when p is small. Several partial data splitting schemes have been proposed as alternatives, the most classical one being V-fold cross-validation (VFCV) [Gei75]. Given some $V \in \{1,\dots,n\}$, VFCV relies on a preliminary partitioning of data into V subsamples of approximately equal cardinality n/V. Each subsample successively plays the role of the validation sample. Formally, let $\mathcal{B}_1,\dots,\mathcal{B}_V$ be some partition of $\{1,\dots,n\}$ with $\mathrm{Card}(\mathcal{B}_j) \approx n/V$ for $j=1,\dots,V$. Then, the VFCV estimator of the risk of $\widehat{s}(D_n)$ is given by Eq. (3.2) with

$$B = V \qquad \text{and} \qquad I_j^{(t)} = \mathcal{B}_j^c \qquad \text{for} \qquad j = 1,\dots,V\,,$$

that is,

$$\widehat{\mathcal{L}}^{\mathrm{VF}}\big(\widehat{s}; D_n; (\mathcal{B}_j)_{1\le j\le V}\big) = \frac{1}{V}\sum_{j=1}^{V} P_n^{(j)}\gamma\Big(\widehat{s}\big(D_n^{(-j)}\big)\Big) \tag{3.4}$$

$$\text{where} \qquad D_n^{(-j)} = (\xi_i)_{i\in\mathcal{B}_j^c} \qquad \text{and} \qquad P_n^{(j)} = \frac{1}{\mathrm{Card}(\mathcal{B}_j)}\sum_{i\in\mathcal{B}_j}\delta_{\xi_i}$$

respectively denote the j-th training sample and the empirical distribution of the j-th validation sample. The computational cost of VFCV is only V times that of training $\widehat{s}$ with $n-n/V$ points, which is much less than LOO or LPO if $V \ll n$. Note that VFCV with $V=n$ is LOO.

3.2. First-order comparison of CV procedures: expectations

Sections 2.2.2–2.2.4 show that for studying CV estimator selection procedures at first order, the key is to compute the expectations of the CV criteria and of $\ell(s^\star, \widehat{s}_m)$ for every $m \in \mathcal{M}$, provided some precise enough concentration inequalities can be obtained. Although proving such concentration inequalities is not straightforward, let us first focus on expectations.

In the general framework of Section 2.1.2, since the splits $(I_j^{(t)})_{1\le j\le B}$ are made independently from the data, we always have

$$\begin{aligned}
&\mathbb{E}\Big[\widehat{\mathcal{L}}^{\mathrm{CV}}\Big(\widehat{s}; D_n; \big(I_j^{(t)}\big)_{1\le j\le B}\Big)\Big] \\
&= \frac{1}{B}\sum_{j=1}^{B}\mathbb{E}\Big[P_n^{(v),j}\gamma\Big(\widehat{s}\big(D_n^{(t),j}\big)\Big)\Big] \\
&= \mathbb{E}\Big[P_n^{(v),1}\gamma\Big(\widehat{s}\big(D_n^{(t),1}\big)\Big)\Big] && \text{since the data are i.i.d. and } \mathrm{Card}\Big(I_j^{(t)}\Big) = n_t \text{ for all } j \\
&= \mathbb{E}\Big[P\gamma\Big(\widehat{s}\big(D_n^{(t),1}\big)\Big)\Big] && \text{since } D_n^{(t),1} \text{ is independent from } D_n^{(v),1} \\
&= \mathbb{E}\Big[P\gamma\big(\widehat{s}(D_{n_t})\big)\Big]
\end{aligned} \tag{3.5}$$

where D_{n_t} is any sample of n_t i.i.d. random variables with common distribution P. In other words, the expectation of the CV criterion is equal to the risk of $\widehat{s}$ trained with n_t data.

So, for estimator selection, it is sufficient to study how the risk of each estimator $\widehat{s}_m$ in the family depends on the sample size for understanding the first-order behavior of all CV criteria.

3.2.1. Bias and suboptimality for estimator selection. Let us assume the following bias-variance decomposition of the risk

$$\forall m \in \mathcal{M}, \quad \mathbb{E}\Big[\ell\big(s^\star, \widehat{s}_m(D_n)\big)\Big] = b(m) + \frac{v(m)}{n} \tag{3.6}$$

for some fonctions $b, v : \mathcal{M} \to \mathbb{R}$. Eq. (3.6) holds true for projection estimators in least-squares density estimation [14]. Eq. (3.6) also approximately holds in particular if Eq. (2.3) holds true with $\alpha_n \propto 1/n$, as for regressograms [16] and for log-likelihood density estimation.

Then, combining Eq. (3.5) and Eq. (3.6), we get

$$\begin{aligned}\mathbb{E}\Big[\widehat{\mathcal{L}}^{\mathrm{CV}}\Big(\widehat{s}; D_n; \big(I_j^{(t)}\big)_{1\le j\le B}\Big)\Big] &= \mathbb{E}\Big[P\gamma\big(\widehat{s}(D_{n_t})\big)\Big] \\ &= P\gamma(s^\star) + b(m) + \frac{v(m)}{n_t} \ . \end{aligned} \tag{3.7}$$

Up to the additive constant $P\gamma(s^\star)$, in expectation, the CV criterion differs from the risk of $\widehat{s}_m$ only by a multiplicative factor n/n_t in front of the "variance" term $v(m)/n$. In particular, for VFCV, this factor is equal to $V/(V-1)$, which stays away from 1 if V is fixed, whereas it tends to 1 as V tends to infinity—as for instance with the LOO ($V = n$) when n tends to infinity.

Since VFCV is biased, it can be suboptimal for estimator selection. More precisely, as proved by Theorem 1 in [16] for regressograms in some specific but realistic setting, if $\mathcal{C}^{\mathrm{VF}}(m)$ denotes the VFCV estimator of the risk of $\widehat{s}_m$ for some partition $\mathcal{B}$ satisfying $\sup_j |\mathrm{Card}(\mathcal{B}_j) - n/V| \le 1$,

$$\begin{aligned}\mathbb{P}\Bigg(\ell\big(s^\star, \widehat{s}_{\widehat{m}_{\mathcal{C}^{\mathrm{VF}}}}(D_n)\big) \ge \Big(1 + \kappa(V) - \log(n)^{-1/5}\Big) \inf_{m\in\mathcal{M}}\Big\{\ell\big(s^\star, \widehat{s}_{\widehat{m}}(D_n)\big)\Big\}\Bigg) \\ \ge 1 - Kn^{-2}\end{aligned} \tag{3.8}$$

for some constants $\kappa(V), K > 0$. In particular, a first-order optimal oracle inequality cannot hold for VFCV, but Eq. (3.8) is a much stronger negative statement: for n large enough, VFCV is (almost) always suboptimal by a factor $\approx 1 + \kappa(V) > 1$.

In order to understand Eq. (3.8) and why we can safely conjecture that it holds much more generally, let us sketch the main ideas behind the proof of Eq. (3.8). Up to concentration inequalities that we discuss in Section 3.2.3, we can reason with expectations. To fix ideas, assume Eq. (3.6) holds with $b(m) = D_m^{-2}$ and $v(m) = \sigma^2 D_m$ for some $\sigma^2, D_m > 0$ such that the set of values taken by the "dimension" D_m is

$\{D_m,\, m \in \mathcal{M}\} = \{1, \dots, n\}$. Then, the risk

$$\mathbb{E}\big[\mathcal{R}(m)\big] = D_m^{-2} + \frac{\sigma^2 D_m}{n}$$

is minimal for $D_m = D_n^\star \approx (2n/\sigma^2)^{1/3}$ and

$$\inf_{m\in\mathcal{M}}\Big\{\mathbb{E}\big[\mathcal{R}(m)\big]\Big\} \approx \Big(2^{-2/3} + 2^{1/3}\Big)\left(\frac{\sigma^2}{n}\right)^{2/3} .$$

By Eq. (3.7), the expectation of the VFCV criterion is

$$\mathbb{E}\big[\mathcal{C}^{\mathrm{VF}}(m)\big] = D_m^{-2} + \frac{V}{V-1}\frac{\sigma^2 D_m}{n}$$

which is minimal for $D_m = \widetilde{D}_n(V) \approx (2C_V n/\sigma^2)^{1/3}$ where $C_V = (V-1)/V$. Therefore,

$$\begin{aligned}
\mathbb{E}\big[\mathcal{R}(\widehat{m}_{\mathcal{C}^{\mathrm{VF}}})\big] \approx \mathbb{E}\Big[\mathcal{R}\big(\widetilde{D}_n(V)\big)\Big] &\approx \Big((2C_V)^{-2/3} + (2C_V)^{1/3}\Big)\left(\frac{\sigma^2}{n}\right)^{2/3} \\
&\approx \frac{(2C_V)^{-2/3} + (2C_V)^{1/3}}{2^{-2/3} + 2^{1/3}} \inf_{m\in\mathcal{M}}\Big\{\mathbb{E}\big[\mathcal{R}(m)\big]\Big\} \\
&= \big(1 + \kappa(V)\big) \inf_{m\in\mathcal{M}}\Big\{\mathbb{E}\big[\mathcal{R}(m)\big]\Big\}
\end{aligned}$$

with $\kappa(V) > 0$ since the function $x \in (0, +\infty) \mapsto (2x)^{-2/3} + (2x)^{1/3}$ admits a unique global minimum at $x = 1 > C_V > 0$. The above arguments can be adapted to any decreasing function $b(m)$ of $v(m)$, if the set of values taken by $b(m)$ and $v(m)$ is "rich enough". Nevertheless, one can clearly build collections of estimators satisfying Eq. (3.6) such that

$$m \mapsto b(m) + \frac{v(m)}{n} \qquad \text{and} \qquad m \mapsto b(m) + \frac{V}{V-1}\frac{v(m)}{n}$$

are minimal for the same $m \in \mathcal{M}$, for instance by considering one "good" estimator and several "very poor" ones. So, even if we cannot state that VFCV is always suboptimal for estimator selection—at first order and when n is large enough—, Eq. (3.8) and its proof suggest that this result holds for most practical problems.

3.2.2. Bias correction and first-order optimal oracle inequalities. The troubles of VFCV can be solved by following the unbiased risk estimation principle presented in Section 2.2.2.

To this aim, Burman [**Bur89**, **Bur90**] proposed a corrected VFCV estimator

$$\widehat{\mathcal{L}}^{\mathrm{corrVF}}(\widehat{s}; D_n) = \widehat{\mathcal{L}}^{\mathrm{VF}}(\widehat{s}; D_n) + P_n\gamma\big(\widehat{s}(D_n)\big) - \frac{1}{V}\sum_{j=1}^{V} P_n\gamma\Big(\widehat{s}\big(D_n^{(-j)}\big)\Big) .$$

Another idea is to use penalization with a resampling-based estimator of the expectation of

$$\mathrm{pen}_{\mathrm{id}}(m; D_n) = (P - P_n)\gamma\big(\widehat{s}_m(D_n)\big)$$

as a penalty, as done in [**Efr83**, **Shi97**] and generalized in **[3]**. For regressograms, **[3]** shows that this leads to an (almost) unbiased estimator of $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m; D_n)]$, and that the resulting estimator selection procedure satisfies a first-order optimal oracle inequality. Using a V-fold subsampling scheme, we get the *V-fold penalty* **[16]**, defined by

$$\mathrm{pen}_{\mathrm{VF}}(\widehat{s}; D_n) = \frac{V-1}{V} \sum_{j=1}^{V} \Big[P_n \gamma\Big(\widehat{s}\big(D_n^{(-j)}\big)\Big) - P_n^{(-j)} \gamma\Big(\widehat{s}\big(D_n^{(-j)}\big)\Big) \Big]$$

where $P_n^{(-j)}$ is the empirical distribution of $D_n^{(-j)}$. It turns out that if $\mathrm{Card}(\mathcal{B}_j) = n/V$ for every $j \in \{1, \ldots, n\}$,

$$P_n - P_n^{(-j)} = \frac{1}{V}\Big(P_n^{(j)} - P_n^{(-j)}\Big) = \frac{1}{V-1}\Big(P_n^{(j)} - P_n\Big)$$

hence

$$\begin{aligned} \mathrm{pen}_{\mathrm{VF}}(\widehat{s}; D_n) &= \frac{V-1}{V^2} \sum_{j=1}^{V} \Big[\big(P_n^{(j)} - P_n^{(-j)}\big) \gamma\Big(\widehat{s}\big(D_n^{(-j)}\big)\Big) \Big] \qquad (3.9) \\ &= \frac{1}{V} \sum_{j=1}^{V} \Big[\big(P_n^{(j)} - P_n\big) \gamma\Big(\widehat{s}\big(D_n^{(-j)}\big)\Big) \Big] \end{aligned}$$

and bias-corrected VFCV from [**Bur89**, **Bur90**] coincides with V-fold penalization:

$$\widehat{\mathcal{L}}^{\mathrm{corrVF}}(\widehat{s}; D_n) = P_n \gamma\big(\widehat{s}(D_n)\big) + \mathrm{pen}_{\mathrm{VF}}(\widehat{s}; D_n) \ .$$

For regressograms **[16]** and for projection estimators in least-squares density estimation **[14]**, we can prove that the V-fold penalty unbiasedly estimates $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m; D_n)]$ and that the corresponding estimator selection procedure satisfies a first-order optimal oracle inequality with a large probability, under mild assumptions. Compared to Eq. (3.8), this shows the advantage of bias-correction, at least at first order.

Note that in the particular case of projection estimators in least-squares density estimation, Lemma 1 in **[14]** shows that VFCV, resampling penalties, LOO and LPO all correspond to penalizing with a V-fold penalty multiplied by some constant $C \geq 1$ (depending on the cross-validation method considered). So, the oracle inequality proved in **[14]** for V-fold penalization also implies that various CV methods satisfy an oracle inequality; for LOO and for LPO with $p \ll n$, this oracle inequality is first-order optimal. We refer to Section 6 of **[7]** for a complete review on oracle inequalities (optimal or not) and other risk bounds available for CV methods.

In order to explain why V-fold penalties should work in an even more general framework—hence be useful for practitioners—, let us sketch the key idea used in **[16]** and **[14]** for showing the V-fold penalty estimates unbiasedly $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m; D_n)]$. Similarly to Eq. (3.6), let us assume that

$$\forall m \in \mathcal{M}, \quad \mathbb{E}\big[\mathrm{pen}_{\mathrm{id}}(m; D_n)\big] = \mathbb{E}\big[(P - P_n)\gamma(\widehat{s}_m)\big] = \frac{w(m)}{n} \qquad (3.10)$$

for some function $w : \mathcal{M} \to \mathbb{R}$, and that $\mathrm{Card}(\mathcal{B}_j) = n/V$ for $j = 1, \dots, V$. Eq. (3.10) holds true in the setting of **[14]** as well as for fixed-design regression with linear estimators (see Sections 4.2 and 4.6), and it holds approximately in the setting of **[16]** as well as for log-likelihood density estimation, at least. By Eq. (3.9),

$$
\begin{aligned}
&\mathbb{E}\big[\mathrm{pen}_{\mathrm{VF}}(\widehat{s}_m; D_n)\big] \\
&= \frac{V-1}{V^2} \sum_{j=1}^{V} \mathbb{E}\Big[\big(P_n^{(j)} - P_n^{(-j)}\big)\gamma\Big(\widehat{s}_m\big(D_n^{(-j)}\big)\Big)\Big] \\
&= \frac{V-1}{V}\, \mathbb{E}\Big[\big(P_n^{(1)} - P_n^{(-1)}\big)\gamma\Big(\widehat{s}_m\big(D_n^{(-1)}\big)\Big)\Big] \qquad \begin{array}{l}\text{since the data are i.i.d.} \\ \text{and } \mathrm{Card}(\mathcal{B}_j) = n/V \text{ for all } j\end{array} \\
&= \frac{V-1}{V}\, \mathbb{E}\Big[\big(P - P_n^{(-1)}\big)\gamma\Big(\widehat{s}_m\big(D_n^{(-1)}\big)\Big)\Big] \qquad \text{since } D_n^{(1)} \text{ is independent from } D_n^{(-1)} \\
&= \frac{V-1}{V}\, \mathbb{E}\Big[\mathrm{pen}_{\mathrm{id}}\big(m; D_n^{(-1)}\big)\Big] \\
&= \frac{V-1}{V} \times \frac{V w(m)}{n(V-1)} = \frac{w(m)}{n} = \mathbb{E}\big[\mathrm{pen}_{\mathrm{id}}(m; D_n)\big] \ .
\end{aligned}
$$

3.2.3. On concentration inequalities. Let us conclude this section by a few comments on what needs to be proved for filling the gap between the above reasonings on expectations and an oracle inequality. As explained in Section 2.2.2, concentration inequalities are needed for $\mathrm{pen}_{\mathrm{VF}}(\widehat{s}_m; D_n)$ and $\mathrm{pen}_{\mathrm{id}}(m; D_n)$ around their expectations. Interestingly, **[16]** and **[14]** propose two different approaches for proving such concentration inequalities, each having pros and cons.

On the one hand, **[16]** suggests a general approach. As soon as a concentration inequality is available for $\mathrm{pen}_{\mathrm{id}}(m; D_n)$, we get a concentration inequality for $\mathrm{pen}_{\mathrm{id}}(m; D_n^{(-j)})$ for every $j = 1, \dots, V$. In addition,

$$\big(P_n^{(j)} - P\big)\gamma\Big(\widehat{s}_m\big(D_n^{(-j)}\big)\Big)$$

is a centered empirical process conditionally to $D_n^{(-j)}$, and standard arguments show it concentrates around its (conditional) expectation, provided n/V is large. Then, according to Eq. (3.9), a union bound over the V folds provides a concentration inequality for $\mathrm{pen}_{\mathrm{VF}}(\widehat{s}_m; D_n)$ around its expectation. The merit of this approach is its generality: one can reasonably conjecture $\mathrm{pen}_{\mathrm{id}}(m; D_n)$ concentrates well around its expectation in many frameworks of practical interest, so this suggests that V-fold penalization satisfies an optimal oracle inequality as soon as Eq. (3.10) approximately holds true. The drawback of these arguments is that the resulting bound only holds if n/V is large enough, that is, V is not too large. For the same reason, the deviation terms increase with V in **[16]**, although empirical observations show that the largest values of V provide the best results.

On the other hand, [14] directly concentrates $\mathrm{pen}_{\mathrm{VF}}(\widehat{s}_m; D_n)$ by writing it as a U-statistic of order two and applying concentration results for such U-statistics [**HRB03**]. Then, we get a bound valid for all V—up to $V = n$, that is, LOO—with remainder terms that do not increase with V, which is closer to empirical observations. However, it is not clear whether this proof can be generalized to other settings since it relies on some specificities of the least-squares density estimation framework.

3.3. Second-order comparison: taking into account the variance

The results presented in Section 3.2 explain how CV methods depend on the size of the training set n_t, but not how they depend on the number B of splits, equal to V for VFCV and V-fold penalization. Nevertheless, it is well known in practice that hold-out performs much worse than VFCV, LOO or LPO, although hold-out with a training set of size

$$n_t = n\left(1 - \frac{1}{\log(n)}\right)$$

is asymptotically optimal and can be proved to satisfy oracle inequalities in many frameworks, similarly to [**vD03, BM06**] for instance. Moreover, simulation experiments in **[16]** and **[14]** clearly show that the estimator selection performance of V-fold penalization improves when V increases.

In order to explain these empirical observations, we must go beyond first-order comparisons, which requires to take into account the "variance" of the CV criteria. Intuitively, the "variance" of hold-out is larger than the "variance" of VFCV or LOO, which is the reason why hold-out performs worse for estimator selection. In order to formalize this intuition, according to the heuristics presented in Section 2.2.5, we must compare the variances of the increments $\mathrm{var}\big(\mathcal{C}(m) - \mathcal{C}(m')\big)$ among several CV criteria $\mathcal{C}$.

For projection estimators in least-squares density estimation, if $\mathcal{C}$ denotes the V-fold penalization criterion with $\mathrm{Card}(\mathcal{B}_j) = n/V$ for $j = 1, \dots, V$, Theorem 2 in **[14]** shows that for every $m, m' \in \mathcal{M}$,

$$\mathrm{var}\big(\mathcal{C}(m) - \mathcal{C}(m')\big) = \left(1 + \frac{4}{V-1} - \frac{1}{n}\right) a_n(m, m') + b_n(m, m') \tag{3.11}$$

for some $a_n(m, m'), b_n(m, m') \geq 0$ which do not depend on V. A similar formula holds for VFCV. Eq. (3.11) is the first result of this form for VFCV methods. Previous variance computations were focusing on the variance of the VFCV criterion—instead of the increments—, and most were asymptotic, see **[7]**.

Eq. (3.11) shows that the variance of V-fold methods actually decreases when V increases—confirming empirical observations—, but the improvement is at most in a second-order term as soon as V is large. Theoretical considerations for nested regular histogram models and simulation experiments suggest that for m, m' that "really matter"

for estimator selection—as explained in Section 2.2.5—,

$$\mathrm{var}\big(\mathcal{C}(m) - \mathcal{C}(m')\big) \propto 1 + \frac{4}{V-1} \ .$$

As a consequence, the improvement from $V = 2$ to 5 or 10 is much larger than from $V = 10$ to $V = n$, which justifies the commonly used principle that taking $V = 5$ or $V = 10$ is good enough [**BS92**, **HTF01**].

3.4. Other works on resampling methods

We conclude this chapter by reporting on some other works on resampling methods, outside the estimator selection setting.

First, in **[2]**, the bootstrap is used for evaluating how well some dynamical models can predict the position of Saturnian satellites outside the observation period, up to two hundred years ahead. After being validated on some simulated data, this procedure has been used for two other problems: (i) estimating quantitatively the improvement of ephemerides induced by the Gaia space mission before it was launched [**Des09**], and (ii) estimating confidence regions for the position of two near-Earth asteroids, in particular for the date and minimal distances of close encouters with Earth [**DAV10**]. The advantage of using resampling methods for these applications is that it provides good confidence regions up to a few hundreds years, for highly non-linear and sensitive models, without strong assumptions on the distribution of errors, contrary to previously used methods for this problem.

Second, in **[10]**, **[4]** and **[5]**, some resampling-based confidence regions and multiple testing procedures are built and theoretically studied, for the mean of high-dimensional Gaussian vectors. Numerical experiments show that these confidence regions and multiple tests automatically adapt to unknown correlation structures between the coordinates of observations.

CHAPTER 4

Minimal penalties

This chapter describes some works on penalization procedures for estimator selection, that is, estimator selection procedures of the form

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ P_n \gamma(\widehat{s}_m) + \operatorname{pen}(m) \right\}. \tag{4.1}$$

More precisely, the following works tackle the notion of minimal penalty: **[1]**, **[11]**, **[18]** and the survey **[22]**.

The initial motivation, presented in Section 4.1, is the practical problem of estimating multiplicative constants in front of penalties. Birgé and Massart [**BM07**] proposed an answer to this problem by considering a related theoretical issue: given a function $\operatorname{pen}_1 : \mathcal{M} \to \mathbb{R}$, how does the penalization procedure defined by Eq. (4.1) with $\operatorname{pen}(m) = C \operatorname{pen}_1(m)$ performs as a function of C? In particular, what is the minimal level of penalization required to get an oracle inequality?

The general heuristics that came out of this work is described in Section 4.3, and made rigorous in Sections 4.4–4.5, in a framework introduced in Section 4.2.

Several empirical results show that this idea can be used fruitfully beyond the framework where it was designed initially [**BMM11**]. A natural question arises: How far can it be safely generalized? Considering linear estimators in regression, **[11]** and **[18]** show that a similar idea can still be used, but at the price of some modification suggested by a theoretical analysis of the problem, as described in Section 4.6.

Finally, as a result of the survey paper **[22]** about minimal penalties from the theoretical point of view, we can summarize in Section 4.7 the settings where minimal penalties are known to work.

4.1. Motivation: Data-driven calibration of constants in front of penalties

Penalties known up to a constant factor appear in several frameworks, for four main reasons:

(1) A penalty satisfying an optimal oracle inequality is theoretically known, but involves *unknown quantities* in practice, such as the residual variance

$$\sigma^2 = \mathbb{E}\left[\varepsilon_i^2 \,|\, X_i\right]$$

for Mallows' C_p and C_L [**Mal73**] in regression (Example 2.2), or the residual covariance matrix in multivariate regression (see **[9]** and Section 6.3).

(2) An optimal penalty pen_1 is known theoretically and in practice, but only *asymptotically*, that is, the (unknown) non-asymptotic optimal penalty is equal to $C_n^\star \mathrm{pen}_1$ with $C_n^\star = 1 + \mathrm{o}(1)$ as the sample size n tends to infinity, but $C_n^\star$ is unknown. For instance, AIC [**Aka73**] and BIC [**Sch78**] penalties for maximum likelihood rely on asymptotic computations.

(3) An optimal penalty is obtained by resampling, hence depending on a multiplicative factor that might depend on unknown quantities or be correct only for n large enough, see **[3]**.

(4) A penalty $C\,\mathrm{pen}_1$ satisfying an oracle inequality with a leading constant $\mathcal{O}(1)$ when C is well-chosen is known theoretically, but theory is not precise enough to specify the optimal value $C^\star$ of C. This occurs for instance for change-point detection [**CR04**, **Leb05**] (see Chapter 5), density estimation with Gaussian mixtures [**MM11**] and local Rademacher complexities in classification [**BBM05**, **Kol06**]. Note that in such cases, it might happen that $C^\star \mathrm{pen}_1$ is not exactly an optimal penalty, so that no oracle inequality with leading constant $1 + \mathrm{o}(1)$ can be obtained; nevertheless, choosing the constant C in the penalty $C\,\mathrm{pen}_1$ remains an important practical problem.

4.2. Fixed-design regression

In order to explain the concept of minimal penalty and how it can be used for calibrating multiplicative constants in front of penalties, we consider the following framework—fixed-design regression—as in [**BM07**], **[11]** and **[18]**.

The framework of Example 2.2 in Section 2.1.3 is modified on two points: (i) the X_i are deterministic, so $(X_i, Y_i)_{1 \leq i \leq n}$ are independent but not identically distributed, and (ii) the excess loss is defined by

$$\frac{1}{n} \sum_{i=1}^{n} \big(t(X_i) - \eta(X_i)\big)^2 \quad \text{where} \quad \forall i \in \{1, \dots, n\}, \quad \eta(X_i) = \mathbb{E}[Y_i] \ . \tag{4.2}$$

As in the random-design regression framework (Example 2.2), we can write that

$$\forall i \in \{1, \dots, n\}, \quad Y_i = \eta(X_i) + \varepsilon_i \quad \text{with} \quad \mathbb{E}[\varepsilon_i] = 0 \quad \text{and} \quad \mathbb{E}\big[\varepsilon_i^2\big] < +\infty \ .$$

We assume that the ε_i are independent and identically distributed with variance σ^2.

Note that we can still define a loss function by $\mathcal{L}(t) = P\gamma(t)$ with γ the least-squares contrast, by defining P as the distribution of (X, Y) where X has a uniform distribution over $\{X_1, \dots, X_n\}$ and $Y = \eta(X) + \varepsilon$ with ε independent from X and distributed as the ε_i. Then, $\mathcal{L}(t)$ is minimal for $t = \eta$ and the excess loss is given by Eq. (4.2).

The fixed-design regression framework is technically easier to analyze because we can cast the regression problem as an estimation problem in $\mathbb{R}^n$ with the Euclidean norm. Let us write

$$Y = (Y_i)_{1 \leq i \leq n} \in \mathbb{R}^n \,, \qquad F = \big(\eta(X_i)\big)_{1 \leq i \leq n} \in \mathbb{R}^n \,, \qquad \varepsilon = (\varepsilon_i)_{1 \leq i \leq n} \in \mathbb{R}^n$$

and consider any $t \in \mathbb{S}$ as an element of $\mathbb{R}^n$ by writing $t_i = t(X_i)$ for $i \in \{1, \dots, n\}$, so that we can assimilate $\mathbb{S}$ with $\mathbb{R}^n$ in the following. Using these notations, we observe

$$Y = F + \varepsilon \in \mathbb{R}^n$$

and the goal is to estimate $s^\star = F$, that is, to find some $t \in \mathbb{R}^n$ such that the excess loss

$$\ell(s^\star, t) = \frac{1}{n}\|t - F\|^2 = \frac{1}{n}\sum_{i=1}^{n}(t_i - F_i)^2$$

is minimal.

Given a family $(S_m)_{m\in\mathcal{M}}$ of linear subspaces of $\mathbb{R}^n$, the models, we consider the estimator selection problem among the family $(\widehat{F}_m)_{m\in\mathcal{M}}$ of corresponding least-squares estimators, which is a model selection problem. Since the empirical risk of t can be written as

$$P_n\gamma(t) = \frac{1}{n}\sum_{i=1}^{n}\big(t(X_i) - Y_i\big)^2 = \frac{1}{n}\|t - Y\|^2 \ ,$$

minimizing it over $t \in S_m$ is equivalent to computing the orthogonal projection $\Pi_m Y$ of Y onto S_m. Hence

$$\forall m \in \mathcal{M}, \quad \widehat{F}_m = \Pi_m Y \ ,$$

a formula that makes explicit computations much simpler than in the random-design case.

4.3. The slope heuristics

Following Section 2.2, the analysis of the estimator selection problem of Section 4.2 starts by computing expectations of the loss and the empirical risk. For every $m \in \mathcal{M}$,

$$\left\|\widehat{F}_m - F\right\|^2 = \left\|(\Pi_m - I_n)F\right\|^2 + \|\Pi_m\varepsilon\|^2 \tag{4.3}$$

and

$$\left\|\widehat{F}_m - Y\right\|^2 = \left\|\widehat{F}_m - F\right\|^2 + \|\varepsilon\|^2 - 2\big\langle\varepsilon,\, \Pi_m\varepsilon\big\rangle + 2\big\langle\varepsilon,\, (I_n - \Pi_m)F\big\rangle \tag{4.4}$$

where

$$\forall t, u \in \mathbb{R}^n, \qquad \langle t,\, u\rangle = \sum_{i=1}^{n} t_i u_i \, .$$

Since the ε_i are independent and centered with variance σ^2, Eq. (4.3) and Eq. (4.4) imply that

$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \frac{1}{n}\left\|(\Pi_m - I_n)F\right\|^2 + \frac{\sigma^2 D_m}{n} \tag{4.5}$$

$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2\right] = \frac{1}{n}\left\|(\Pi_m - I_n)F\right\|^2 + \sigma^2 - \frac{\sigma^2 D_m}{n} \tag{4.6}$$

where $D_m := \dim(S_m)$. Therefore, the expectation of the ideal penalty is

$$\mathbb{E}\big[\mathrm{pen}_{\mathrm{id}}(m)\big] = \mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 - \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2\right] = \frac{2\sigma^2 D_m}{n} - \sigma^2$$

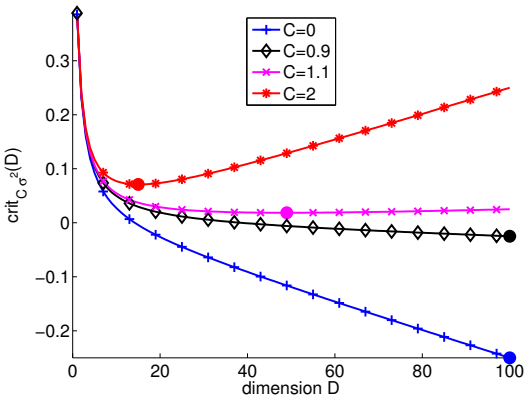


FIGURE 1. Plot of $\mathrm{crit}_{C\sigma^2}(m)$ as a function of D_m for $C \in \{0, 0.9, 1.1, 2\}$; its minimal value at $m^\star(C\sigma^2)$ is shown by a solid circle. Same setting as Figure 1.

which provides the following (first-order) optimal penalty

$$\mathrm{pen}_{\mathrm{opt}}(m) = \frac{2\sigma^2 D_m}{n} \ , \tag{4.7}$$

since adding a constant σ^2 to the penalty does not change the selected model. The penalty $\mathrm{pen}_{\mathrm{opt}}$ is called Mallows' C_p [**Mal73**].

Eq. (4.7) shows that the shape $\mathrm{pen}_1(m) = D_m/n$ of the optimal penalty is known, even when σ^2 is unknown. At this point it is natural to ask what is the minimal value of the constant that should be put in front of $\mathrm{pen}_1(m)$. More precisely, if for every $C \geq 0$,

$$\widehat{m}(C) \in \operatorname*{argmin}_{m\in\mathcal{M}} \left\{ \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + C\frac{D_m}{n} \right\} \ , \tag{4.8}$$

what is the minimal value of C such that $\widehat{m}(C)$ is a "reasonable" choice, that is, avoids strong overfitting, or equivalently, satisfies an oracle inequality (2.2) with $C_n = \mathcal{O}(1)$ as n tends to infinity?

In order to understand how $\widehat{m}(C)$ behaves as a function of C, let us consider, for every $C \geq 0$,

$$m^\star(C) \in \operatorname*{argmin}_{m\in\mathcal{M}} \left\{ \mathbb{E}\left[\frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + C\frac{D_m}{n} \right] \right\} = \operatorname*{argmin}_{m\in\mathcal{M}} \left\{ \mathrm{crit}_C(m) \right\}$$
$$\text{with} \quad \mathrm{crit}_C(m) := \frac{1}{n} \left(\left\| (\Pi_m - I_n)F \right\|^2 + (C - \sigma^2)D_m \right) \ , \tag{4.9}$$

where the above equality comes from Eq. (4.6). Provided we can prove some uniform concentration inequalities for $\|\widehat{F}_m - Y\|^2$, we can expect $m^\star(C)$ to be close to $\widehat{m}(C)$. Let us also assume, for simplicity, that the approximation error $n^{-1}\|(\Pi_m - I_n)F\|^2$ is a decreasing function of D_m—which holds for instance if the S_m are nested—and is almost constant for D_m large enough. Then, two cases can be distinguished with respect to C :

– if $C < \sigma^2$, then $\mathrm{crit}_C(m)$ is a decreasing function of D_m, so $D_{m^\star(C)}$ is huge: $m^\star(C)$ overfits.
– if $C > \sigma^2$, then $\mathrm{crit}_C(m)$ increases with D_m for D_m large enough, so $D_{m^\star(C)}$ is much smaller than when $C < \sigma^2$.

This behaviour is illustrated on Figure 1. In other words, $\sigma^2 \,\mathrm{pen}_1(m)$ seems to be the minimal amount of penalization needed so that a minimizer $\widehat{m}$ of the penalized criterion does not clearly overfit. The above arguments are made rigorous in Section 4.4, showing that

$$\mathrm{pen}_{\min}(m) := \frac{\sigma^2 D_m}{n} \tag{4.10}$$

is indeed a minimal penalty in the current framework.

We can now summarize Birgé and Massart's slope heuristics [**BM07**] into two major facts. First, from Eq. (4.7) and (4.10), we get a *relationship between the optimal and minimal penalties*:

$$\mathrm{pen}_{\mathrm{opt}}(m) = 2\,\mathrm{pen}_{\min}(m) \ . \tag{4.11}$$

Second, *the minimal penalty is observable*, since $D_{\widehat{m}(C)}$ decreases "smoothly" as a function of C everywhere except around $C = \sigma^2$ where it jumps.

Data-driven calibration algorithm. The two major facts of the slope heuristics described above directly lead to a data-driven penalty algorithm: we can estimate the minimal penalty by looking for a jump of $D_{\widehat{m}(C)}$ as a function of C, and make use of Eq. (4.11) to get an estimator of the optimal penalty.

Algorithm 1.
`Input`: $\left(\|\widehat{F}_m - Y\|^2\right)_{m\in\mathcal{M}}$.

(1) Compute $\left(\widehat{m}(C)\right)_{C\geq 0}$, where $\widehat{m}(C)$ is defined by Eq. (4.8).

(2) Find $\widehat{C}_{\mathrm{jump}} > 0$ corresponding to the "unique large jump" of $C \mapsto D_{\widehat{m}(C)}$.

(3) Select

$$\widehat{m}_{\mathrm{Alg.1}} \in \operatorname*{argmin}_{m\in\mathcal{M}} \left\{ n^{-1} \left\| \widehat{F}_m - Y \right\|^2 + \frac{2\widehat{C}_{\mathrm{jump}} D_m}{n} \right\}.$$

`Output`: $\widehat{m}_{\mathrm{Alg.1}}$.

Figure 2 shows a plot of $C \mapsto D_{\widehat{m}(C)}$ for one sample, with one clear jump corresponding to $\widehat{C}_{\mathrm{jump}}$. Computational and practical issues related to Algorithm 1—in particular, how to define properly $\widehat{C}_{\mathrm{jump}}$—are discussed in **[22]**. Algorithm 1 can also be related to some "L-curve", "corner" or "elbow" heuristics which are often used in machine learning [**HO93**], see **[22]**.

4.4. Theoretical result for least-squares regression

The heuristics of Section 4.3 can be formalized with the following theorem.

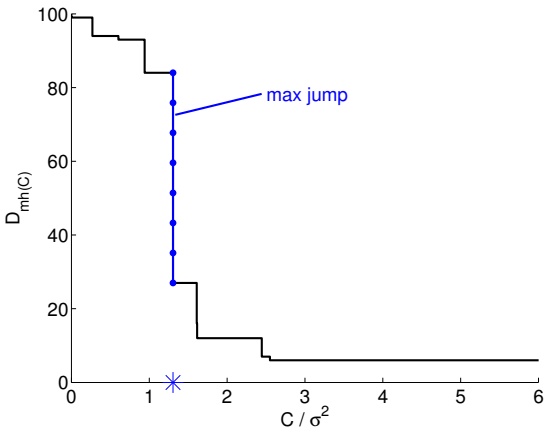


FIGURE 2. Illustration of Algorithm 1 on one sample: Plot of the function $C \mapsto D_{\widehat{m}(C)}$ and visualization of $\widehat{C}_{\text{jump}}$. Same setting as Figure 1.

Theorem 4.1 (Theorem 1 in **[22]**). *In the framework described in Section 4.2, assume that $\mathcal{M}$ is finite and that*

$$\exists m_1 \in \mathcal{M}\,, \quad S_{m_1} = \mathbb{R}^n \ , \tag{HId}$$

$$\inf_{m\in\mathcal{M}} \left\{ \mathbb{E}\left[\frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right] \right\} \leq \sigma^2 \delta_n \qquad \text{with} \quad \delta_n \leq \frac{1}{20} \ , \tag{HO}$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n) \ . \tag{HG}$$

Recall that for every $C \geq 0$, $\widehat{m}(C)$ is defined by Eq. (4.8). Then, for every $\gamma \geq 0$, some $n_0(\gamma)$ exists such that if $n \geq n_0(\gamma)$, with probability at least $1 - 4\,\mathrm{Card}(\mathcal{M}) n^{-\gamma}$, the following holds simultaneously:

$$\forall C < (1 - \eta_n^-)\sigma^2\,, \qquad D_{\widehat{m}(C)} \geq \frac{9n}{10} \tag{4.12}$$

$$\forall C > (1 + \eta_n^+)\sigma^2\,, \qquad D_{\widehat{m}(C)} \leq \frac{n}{10} \ , \tag{4.13}$$

and for every $\eta \in (0, 1/2]$ and $C \in \left[(2-\eta)\sigma^2, (2+\eta)\sigma^2\right]$,

$$\frac{1}{n}\left\| \widehat{F}_{\widehat{m}(C)} - F \right\|^2 \leq (1+3\eta) \inf_{m\in\mathcal{M}} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right\} + \frac{880\sigma^2\gamma\log(n)}{\eta n} \tag{4.14}$$

$$\text{where} \qquad \eta_n^- = 81\sqrt{\frac{\gamma\log(n)}{n}} \quad \text{and} \quad \eta_n^+ = 20\delta_n + 81\sqrt{\frac{\gamma\log(n)}{n}} \ .$$

Theorem 4.1—which comes from **[22]**—is proved in Section 4.5. It revisits a result by [**BM07**] by using some arguments from **[18]**, under milder assumptions than the ones of **[18]**.

[**BM07**] is the first article where $\mathrm{pen}_{\min}$ is proved to be a minimal penalty, but with a weaker statement compared to Theorem 4.1. In order to compare the two statements, let us assume that $\widehat{C}_{\mathrm{jump}}$ in Algorithm 1 is defined by

$$\inf\left\{C > 0 \,/\, D_{\widehat{m}(C)} \leq n/2\right\} \ .$$

Then, a novelty of Theorem 4.1 is to prove that $\widehat{C}_{\mathrm{jump}}$ estimates consistently σ^2 as soon as $\delta_n = \mathrm{o}(1)$ in assumption (**HO**), that is, if the oracle model consistently estimates the signal F. Actually, Theorem 4.1 is even more precise since it gives a non-asymptotic upper bound on the difference between $\widehat{C}_{\mathrm{jump}}$ and σ^2. Eq. (4.12) and (4.13) also imply corresponding lower/upper bounds on the risk of $\widehat{F}_{\widehat{m}(C)}$, confirming that $\mathrm{pen}_{\min}$ also is a minimal penalty in terms of risk, as stated in the heuristics of Section 4.3.

Finally, we comment on the assumptions of Theorem 4.1: (**HId**) can always be satisfied, and (**HG**) could be replaced by any other noise assumption such that the two concentration inequalities (4.15)–(4.16) used in the proof of Theorem 4.1 are satisfied. Theorem 4.1 also implicitly assumes $\mathrm{Card}(\mathcal{M}) \ll n^{-\gamma}$ for some $\gamma > 0$—otherwise, the result does not hold on a large probability event—, but such an assumption is classical and somehow unavoidable for procedures relying on the unbiased risk estimation principle, see Section 2.2.2. Overall, the strongest assumption in Theorem 4.1 is (**HO**), and it is likely to be minimal: We cannot hope to estimate consistently the residual variance σ^2 from a family of models $(S_m)_{m\in\mathcal{M}_n}$ which cannot consistently estimate the signal F. Note that many other approaches for estimating σ^2 exist in the literature, under similar or stronger assumptions, see **[22]**.

4.5. Proof of Theorem 4.1

We report here the proof of Theorem 4.1, as done in **[22]**, since it illustrates well the general approach presented in Section 2.2 while being simple enough technically: fixed-design regression with least-squares estimators is probably the simpler setting for proving such a result. Theorem 4.1 actually contains two main types of results: Eq. (4.14) is an optimal oracle inequality and Eq. (4.12) provides a (rough) lower bound on the performance of some model selection procedure.

The proof mixes ideas from [**BM07**] and **[18]**. We split it into three main steps, the last two ones being split itself into several substeps: (1) using concentration inequalities, (2) proving the existence of a dimension jump (Eq. (4.12)–(4.13)), (3) proving an oracle inequality (Eq. (4.14)).

Step 1: concentration inequalities. The slope heuristics presented in Section 4.3 relies on the fact that $\|\widehat{F}_m - Y\|^2$ and $\|\widehat{F}_m - F\|^2$ are close to their respective expectations. Given Eq. (4.3)–(4.4), for every $m \in \mathcal{M}$ and $x \geq 0$, we consider the event $\Omega_{m,x}$ on which the following two inequalities hold simultaneously:

$$\left|\langle \varepsilon,\, \Pi_m \varepsilon\rangle - \sigma^2 D_m\right| \leq 2\sigma^2\sqrt{x D_m} + 2x\sigma^2 \tag{4.15}$$

$$\left|\langle \varepsilon,\, (I_n - \Pi_m)F\rangle\right| \leq \sigma\sqrt{2x}\,\left\|(I_n - \Pi_m)F\right\| \ . \tag{4.16}$$

Under (**HG**), by standard Gaussian concentration results, for instance Propositions 4 and 6 in [18], we have

$$\mathbb{P}(\Omega_{m,x}) \geq 1 - 4e^{-x} \ .$$

Then, defining $\Omega_x := \bigcap_{m\in\mathcal{M}} \Omega_{m,x}$, the union bound gives

$$\mathbb{P}(\Omega_x) \geq 1 - 4\,\mathrm{Card}(\mathcal{M})e^{-x}$$

and it is sufficient to prove that Eq. (4.12)–(4.14) hold true on Ω_x with $x = \gamma\log(n)$.

From now on, we assume $x = \gamma\log(n)$ and $x/n = \gamma\log(n)/n \leq 1/81^2$ (which defines $n_0(\gamma)$), and we restrict ourselves to the event Ω_x. At various places in the proof (in steps 2.2, 2.3, 3.2, 3.3), we make use of the inequality $2\sqrt{ab} \leq \theta a + \theta^{-1}b$ for all $a, b, \theta > 0$.

Step 2: existence of a dimension jump. Let us define

$$m_2 \in \operatorname*{argmin}_{m\in\mathcal{M}}\left\{\mathbb{E}\left[\left\|\widehat{F}_m - F\right\|^2\right]\right\}.$$

For proving Eq. (4.12) and (4.13), we show that $\widehat{m}(C)$ minimizes a quantity $G_C(m)$ close to $\mathrm{crit}_C(m)$, and then we show that $G_C(m_1)$ (resp. $G_C(m_2)$) is smaller than $G_C(m)$ for any model m with $D_m < 9n/10$ (resp. $D_m > n/10$).

Step 2.1: control of the difference between $\mathrm{crit}_C(m)$ *and* $G_C(m)$. Let $C \geq 0$. By Eq. (4.8), (4.3) and (4.4), since $\|\varepsilon\|^2$ does not depend from m, we get that $\widehat{m}(C)$ minimizes

$$\begin{aligned} G_C(m) &:= \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + C\frac{D_m}{n} - \frac{1}{n}\|\varepsilon\|^2 \\ &= \frac{1}{n}\left\|(I_n - \Pi_m)F\right\|^2 - \frac{1}{n}\langle\varepsilon,\, \Pi_m\varepsilon\rangle + C\frac{D_m}{n} + \frac{2}{n}\langle\varepsilon,\, (I_n - \Pi_m)F\rangle \\ &= \frac{1}{n}\,\mathrm{crit}_C(m) - \left(\frac{1}{n}\langle\varepsilon,\, \Pi_m\varepsilon\rangle - \sigma^2 D_m\right) + \frac{2}{n}\langle\varepsilon,\, (I_n - \Pi_m)F\rangle \end{aligned}$$

where crit_C is defined by Eq. (4.9). Therefore, by Eq. (4.15)–(4.16) and using that $D_m \leq n$, for every $m \in \mathcal{M}$,

$$\left|G_C(m) - \frac{1}{n}\,\mathrm{crit}_C(m)\right| \leq 4\sigma^2\sqrt{\frac{x}{n}} + \frac{2\sigma\sqrt{2x}}{n}\left\|(I_n - \Pi_m)F\right\| \ . \tag{4.17}$$

Step 2.2: lower bound on $D_{\widehat{m}(C)}$ *when* C *is too small (proof of Eq. (4.12)).* Since for every $C \geq 0$, $\widehat{m}(C)$ minimizes $G_C(m)$ over $m \in \mathcal{M}$, it is sufficient to prove that for $C \in [0, \sigma^2)$ far enough from σ^2,

$$G_C(m_1) < \inf_{m\in\mathcal{M},\, D_m < 9n/10}\left\{G_C(m)\right\}$$

where m_1 is given by (**HId**). Let $C \in [0, \sigma^2)$. On the one hand, by Eq. (4.17),

$$G_C(m_1) \leq \frac{1}{n}\,\mathrm{crit}_C(m_1) + 4\sigma^2\sqrt{\frac{x}{n}} = C - \sigma^2 + 4\sigma^2\sqrt{\frac{x}{n}} \ . \tag{4.18}$$

On the other hand, by Eq. (4.17), for any $m \in \mathcal{M}$ such that $D_m < 9n/10$,

$$G_C(m) \geq \frac{(C-\sigma^2)D_m}{n} - 4\sigma^2\sqrt{\frac{x}{n}} + \frac{1}{n}\left\|(I_n - \Pi_m)F\right\|^2 - \frac{2\sigma\sqrt{2x}}{n}\left\|(I_n - \Pi_m)F\right\|$$
$$\geq \frac{9}{10}(C-\sigma^2) - 4.05\sigma^2\sqrt{\frac{x}{n}} \ . \tag{4.19}$$

To conclude, the upper bound in Eq. (4.18) is strictly smaller than the lower bound in Eq. (4.19) if and only if

$$C < \sigma^2\left(1 - 81\sqrt{\frac{\gamma\log(n)}{n}}\right) \ .$$

Step 2.3: *upper bound on $D_{\widehat{m}(C)}$ when C is large enough* (*proof of Eq.* (4.13)). Similarly to the proof of Eq. (4.12), it is sufficient to prove that for $C > \sigma^2$ far enough from σ^2,

$$G_C(m_2) < \inf_{m\in\mathcal{M}\,,\, D_m > n/10}\left\{G_C(m)\right\}$$

Let $C > \sigma^2$. On the one hand, by Eq. (4.17) and (**HO**),

$$G_C(m_2) \leq \frac{1}{n}\operatorname{crit}_C(m_2) + 4\sigma^2\sqrt{\frac{x}{n}} + \frac{2\sigma\sqrt{2x}}{n}\left\|(I_n - \Pi_{m_2})F\right\|$$
$$\leq \frac{2}{n}\left\|(I_n - \Pi_{m_2})F\right\|^2 + \frac{(C-\sigma^2)D_{m_2}}{n} + 4\sigma^2\sqrt{\frac{x}{n}} + \frac{2\sigma^2 x}{n}$$
$$\leq \sigma^2\left(\max\left\{2, \frac{C}{\sigma^2} - 1\right\}\delta_n + 4.05\sqrt{\frac{x}{n}}\right) \ . \tag{4.20}$$

On the other hand, by Eq. (4.17), for any $m \in \mathcal{M}$ such that $D_m > n/10$,

$$G_C(m) \geq \frac{1}{10}(C-\sigma^2) - 4.05\sigma^2\sqrt{\frac{x}{n}} \ . \tag{4.21}$$

To conclude, the upper bound in Eq. (4.20) is strictly smaller than the lower bound in Eq. (4.21) if $x/n \leq 1/81^2$ and

$$\frac{C}{\sigma^2} - 1 > 20\delta_n + 81\sqrt{\frac{x}{n}} =: \eta_n^+ \ .$$

Step 3: oracle inequality. For proving Eq. (4.14), we prove a slightly more general oracle inequality—Eq. (4.24)—using the classical approach described in Section 2.2.1.

Step 3.1: In order to apply Lemma 2.1, let us define

$$\mathcal{R}(m) = \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 \quad \text{and} \quad \mathcal{C}(m) = \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + \frac{CD_m}{n} - \|\varepsilon\|^2 \ .$$

Then, by Eq. (4.4), for every $m \in \mathcal{M}$,

$$\mathcal{C}(m) - \mathcal{R}(m) = \frac{(C-2\sigma^2)D_m}{n} - \Delta(m)$$

$$\text{with} \qquad \Delta(m) := \frac{2}{n}\big(\langle \varepsilon,\, \Pi_m \varepsilon\rangle - \sigma^2 D_m\big) - \frac{2}{n}\big\langle \varepsilon,\, (I_n - \Pi_m)F\big\rangle \ .$$

In order to find some $A(m)$ and $B(m)$ such that Eq. (2.8) holds true on Ω_x, we need to control $\Delta(m)$.

Step 3.2: *control of* $\Delta(m)$. By Eq. (4.15) and (4.16), on Ω_x, for every $m \in \mathcal{M}$ and $\theta > 0$,

$$\begin{aligned} |\Delta(m)| &\le \frac{2}{n}\Big[2\sigma^2\sqrt{xD_m} + 2\sigma^2 x + \sigma\sqrt{2x}\big\|(I_n - \Pi_m)F\big\|\Big] \\ &\le 2\theta\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] + \frac{\sigma^2 x}{n}\big(3\theta^{-1} + 4\big) \ . \end{aligned} \tag{4.22}$$

Step 3.3: *upper bound on the expected loss in terms of loss.* By Eq. (4.3) and (4.16), on Ω_x, for every $m \in \mathcal{M}$ and $\theta' > 0$,

$$\begin{aligned} \left\|\widehat{F}_m - F\right\|^2 &= \mathbb{E}\left[\left\|\widehat{F}_m - F\right\|^2\right] + \langle \varepsilon,\, \Pi_m\varepsilon\rangle - \sigma^2 D_m \\ &\ge \mathbb{E}\left[\left\|\widehat{F}_m - F\right\|^2\right] - \sigma^2\Big(2\sqrt{xD_m} + 2x\Big) \\ &\ge (1-\theta')\mathbb{E}\left[\left\|\widehat{F}_m - F\right\|^2\right] - x\sigma^2\big(2 + \theta'^{-1}\big) \end{aligned}$$

so that, for every $\theta' \in (0,1)$,

$$\mathbb{E}\left[\left\|\widehat{F}_m - F\right\|^2\right] \le \frac{1}{1-\theta'}\left\|\widehat{F}_m - F\right\|^2 + \kappa(\theta')x\sigma^2 \qquad \text{with} \quad \kappa(\theta') := \frac{2 + \frac{1}{\theta'}}{1-\theta'} \ . \tag{4.23}$$

Step 3.4: *definition of* $A(m)$ *and* $B(m)$. Combining Eq. (4.22) and (4.23), we get on the one hand that, on Ω_x, for every $m \in \mathcal{M}$, $\theta > 0$, $\theta' \in (0,1)$,

$$\begin{aligned} &\Delta(m) + \frac{\big(2\sigma^2 - C\big)D_m}{n} \\ &\le \left(2\theta + \left(2 - \frac{C}{\sigma^2}\right)_+\right)\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] + \frac{\sigma^2 x}{n}\big(3\theta^{-1} + 4\big) \\ &\le A(m) := \frac{2\theta + \big(2 - \frac{C}{\sigma^2}\big)_+}{1-\theta'}\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 + \frac{\sigma^2 x}{n}\left(\frac{3}{\theta} + 4 + \kappa(\theta')\left[2\theta + \left(2 - \frac{C}{\sigma^2}\right)_+\right]\right) \ . \end{aligned}$$

On the other hand, similarly, on Ω_x, for every $m \in \mathcal{M}$, $\theta > 0$, $\theta' \in (0,1)$,

$$\begin{aligned} &-\Delta(m) + \frac{\big(C - 2\sigma^2\big)D_m}{n} \\ &\le B(m) := \frac{2\theta + \big(\frac{C}{\sigma^2} - 2\big)_+}{1-\theta'}\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 + \frac{\sigma^2 x}{n}\left(\frac{3}{\theta} + 4 + \kappa(\theta')\left[2\theta + \left(\frac{C}{\sigma^2} - 2\right)_+\right]\right) \ . \end{aligned}$$

Step 3.5: *proof of a general oracle inequality.* Applying Lemma 2.1 with $\mathcal{C}$ and $\mathcal{R}$ defined as in step 3.1, A and B defined as in step 3.4, and $\theta' = 2\theta$, we get that on Ω_x,

for every $\theta \in (0,1/2)$ and $m \in \mathcal{M}$,

$$\left(1 - \frac{2\theta + \left(2 - \frac{C}{\sigma^2}\right)_+}{1-2\theta}\right)\frac{1}{n}\left\|\widehat{F}_{\widehat{m}(C)} - F\right\|^2 \leq \left(1 + \frac{2\theta + \left(\frac{C}{\sigma^2} - 2\right)_+}{1-2\theta}\right)\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 + \frac{\sigma^2 x}{n} R_1\left(\theta, C\sigma^{-2}\right)$$

$$\text{with} \qquad R_1\left(\theta, C\sigma^{-2}\right) := \frac{6}{\theta} + 8 + \kappa(2\theta)\left(4\theta + \left|\frac{C}{\sigma^2} - 2\right|\right) .$$

Now, for any $\delta \in (0,1]$, we choose

$$\theta = \theta^\star(\delta) := \frac{\delta}{4}\min\left\{\frac{1}{1+\delta+\left(\frac{C}{\sigma^2}-2\right)_+}, \frac{\left(1-\left(2-\frac{C}{\sigma^2}\right)_+\right)^2}{1+\delta\left(1-\left(2-\frac{C}{\sigma^2}\right)_+\right)}\right\} \leq \frac{1}{4} ,$$

so that we get, if $C \geq (1+\delta)\sigma^2$,

$$\frac{1}{n}\left\|\widehat{F}_{\widehat{m}(C)} - F\right\|^2 \leq \left(\frac{1 + \left(\frac{C}{\sigma^2} - 2\right)_+}{1 - \left(2 - \frac{C}{\sigma^2}\right)_+} + \delta\right) \times \left(\inf_{m\in\mathcal{M}}\left\{\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right\} + \frac{\sigma^2 x}{n} R_2\left(\delta, \frac{C}{\sigma^2}\right)\right) \tag{4.24}$$

where, for every $\delta \in (0,1]$ and $u \in (1,+\infty)$,

$$R_2(\delta, u) \leq \frac{8}{\delta}\left(5 + |u-2|\right)\max\left\{2 + (u-2)_+ , \frac{2}{\left(1-(2-u)_+\right)^2}\right\} .$$

Step 3.6: oracle inequality for $\widehat{m}(C)$ when C is close to $2\sigma^2$ (proof of Eq. (4.14)). Now, we assume $C/\sigma^2 \in [2-\eta, 2+\eta]$ with $\eta \in (0,1/2]$. Taking $\delta = \eta$ in Eq. (4.24) yields

$$\begin{aligned}\frac{1}{n}\left\|\widehat{F}_{\widehat{m}(C)} - F\right\|^2 &\leq \left(\max\left\{1+\eta, \frac{1}{1-\eta}\right\} + \delta\right)\left(\inf_{m\in\mathcal{M}}\left\{\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right\} + \frac{\sigma^2 x}{n}\frac{8}{\delta}(5+\eta)\max\left\{2+\eta, \frac{2}{(1-\eta)^2}\right\}\right) \\ &\leq (1+3\eta)\inf_{m\in\mathcal{M}}\left\{\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right\} + \frac{880\sigma^2 x}{\eta n}\end{aligned}$$

where we used the fact that $1/(1-\eta) \leq 1+2\eta$ for every $\eta \in [0,1/2]$. □

4.6. Extension to linear estimators: minimal penalties

The slope heuristics empirically works well in several cases outside the framework of Theorem 4.1 [**BMM11**]. So, a natural question arises: how far can it be generalized? Given the formulation (4.11) of the slope heuristics, if pen_1 is the shape of an optimal penalty and if $\mathbb{C}(m)$ is a complexity measure associated with $\widehat{s}_m$—large when $\widehat{s}_m$ overfits,

small when $\widehat{s}_m$ underfits—, the natural generalization of Algorithm 1 is the following, with the notation of Section 2.1.2.

Algorithm 2.
Input: $\big(P_n\gamma(\widehat{s}_m)\big)_{m\in\mathcal{M}}$, $\big(\mathrm{pen}_1(m)\big)_{m\in\mathcal{M}}$ and $(\mathbb{C}_m)_{m\in\mathcal{M}}$.

(1) Compute $\big(\widehat{m}_1(C)\big)_{C\geq 0}$, where for every $C \geq 0$,

$$\widehat{m}_1(C) \in \operatorname*{argmin}_{m\in\mathcal{M}} \big\{ P_n\gamma(\widehat{s}_m) + C\,\mathrm{pen}_1(m) \big\} \ . \tag{4.25}$$

(2) Find $\widehat{C}_{\mathrm{jump}} > 0$ corresponding to the "unique large jump" of $C \mapsto \mathbb{C}_{\widehat{m}_1(C)}$.

(3) Select

$$\widehat{m}_{\text{Alg.2}} \in \operatorname*{argmin}_{m\in\mathcal{M}} \Big\{ P_n\gamma(\widehat{s}_m) + 2\widehat{C}_{\mathrm{jump}}\,\mathrm{pen}_1(m) \Big\} .$$

Output: $\widehat{m}_{\text{Alg.2}}$.

In order to test the general validity of Algorithm 2, let us consider the problem of selecting among linear estimators in fixed-design regression, as in **[11]** and **[18]**. Compared to the model selection problem described in Section 4.2, the only difference is that

$$\forall m \in \mathcal{M}, \quad \widehat{F}_m = A_m Y$$

for some deterministic $n \times n$ matrix A_m, without requiring A_m to be an orthogonal projection matrix. Classical examples of linear estimators are projection (least-squares) estimators, kernel ridge regression [**SS01**]—also known as spline smoothing when using spline kernels [**Wah90**]—, nearest-neighbor regression and Nadaraya-Watson estimators [**Nad64**, **Wat64**]; see **[18]** for more examples and references. More details on kernel ridge regression can be found in Section 6.3.

As in Section 4.3, expectations of the loss and the empirical risk of a linear estimator can be computed as follows:

$$\Big\|\widehat{F}_m - F\Big\|^2 = \big\|(A_m - I_n)F\big\|^2 + \|A_m\varepsilon\|^2 + 2\big\langle A_m\varepsilon,\, (A_m - I_n)F\big\rangle \ , \tag{4.26}$$

$$\Big\|\widehat{F}_m - Y\Big\|^2 = \Big\|\widehat{F}_m - F\Big\|^2 + \|\varepsilon\|^2 - 2\langle\varepsilon,\, A_m\varepsilon\rangle + 2\big\langle\varepsilon,\, (I_n - A_m)F\big\rangle \ , \tag{4.27}$$

so that

$$\mathbb{E}\left[\frac{1}{n}\Big\|\widehat{F}_m - F\Big\|^2\right] = \frac{1}{n}\big\|(A_m - I_n)F\big\|^2 + \frac{\sigma^2\,\mathrm{tr}\big(A_m^\top A_m\big)}{n} \tag{4.28}$$

$$\mathbb{E}\left[\frac{1}{n}\Big\|\widehat{F}_m - Y\Big\|^2\right] = \frac{1}{n}\big\|(A_m - I_n)F\big\|^2 + \sigma^2 + \frac{\sigma^2\big(\mathrm{tr}\big(A_m^\top A_m\big) - 2\,\mathrm{tr}(A_m)\big)}{n} . \tag{4.29}$$

We deduce that

$$\mathrm{pen}_{\mathrm{opt}}(m) = \mathbb{E}\left[\frac{1}{n}\Big\|\widehat{F}_m - F\Big\|^2 - \frac{1}{n}\Big\|\widehat{F}_m - Y\Big\|^2\right] + \sigma^2 = \frac{2\sigma^2\,\mathrm{tr}(A_m)}{n} \tag{4.30}$$

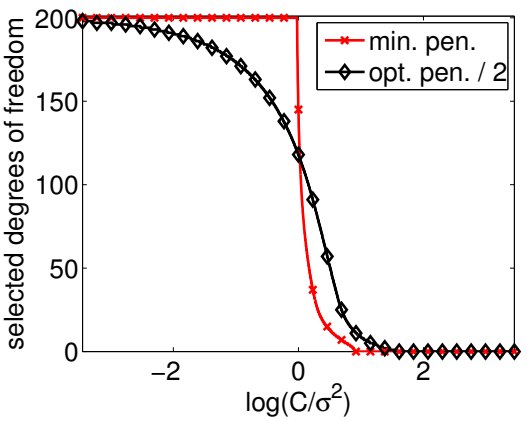


Figure 3. The minimal penalty is not proportional to $\mathrm{tr}(A_m)$ for kernel ridge estimators (figure taken from [18], 'kernel ridge' framework): plots of $C \mapsto \mathbb{C}_{\widehat{m}_1(C)}$ for Algorithm 2 with $\mathrm{pen}_1(m) = \mathrm{tr}(A_m)/n$ and $\mathbb{C}_m = \mathrm{tr}(A_m)$ (black diamonds), and $C \mapsto \mathbb{C}_{\widehat{m}_{\min}(C)}$ for Algorithm 3 with $\mathrm{pen}_0(m) = (2\,\mathrm{tr}(A_m) - \mathrm{tr}(A_m^\top A_m))/n$ and $\mathbb{C}_m = \mathrm{tr}(A_m)$ (red crosses).

is an optimal penalty, called Mallows' C_L [Mal73].

As for projection estimators, the C_L penalty depends on σ^2, which must be estimated. Since C_L corresponds to C_p with the dimension D_m replaced by the degrees of freedom $\mathrm{tr}(A_m)$, a natural idea is to apply Algorithm 2 with

$$\widehat{s}_m = \widehat{F}_m\,, \quad P_n\gamma(\widehat{s}_m) = \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2, \quad \mathrm{pen}_1(m) = \frac{\mathrm{tr}(A_m)}{n} \quad \text{and} \quad \mathbb{C}_m = \mathrm{tr}(A_m)\,.$$

Simulation experiments in [18], reported on Figure 3, clearly show this fails: no clear jump of $\mathbb{C}_{\widehat{m}_1(C)}$ can be observed as a function of C!

How can we fix this failure? Let us come back to our computations of expectations, that is, Eq. (4.28)–(4.29). Generally, the approximation error $n^{-1}\|(A_m - I_n)F\|^2$ is almost constant when $\mathrm{tr}(A_m)$ is large, and $2\,\mathrm{tr}(A_m) - \mathrm{tr}\big(A_m^\top A_m\big)$ is a positive increasing function of $\mathrm{tr}(A_m)$. Therefore, adding

$$C \times \frac{2\,\mathrm{tr}(A_m) - \mathrm{tr}\big(A_m^\top A_m\big)}{n}$$

to the expectation of the empirical risk yields (approximately) a decreasing function of $\mathrm{tr}(A_m)$ if $C < \sigma^2$, and—for large values of $\mathrm{tr}(A_m)$—an increasing function of $\mathrm{tr}(A_m)$ if $C > \sigma^2$. In other words,

$$\mathrm{pen}^{\mathrm{lin}}_{\min}(m) = \sigma^2\,\frac{2\,\mathrm{tr}(A_m) - \mathrm{tr}\big(A_m^\top A_m\big)}{n} \tag{4.31}$$

is a minimal penalty for our estimator selection problem. Since $\mathrm{pen}_{\min}^{\mathrm{lin}}$ is known up to the multiplicative factor σ^2, exactly as $\mathrm{pen}_{\mathrm{opt}}$, we propose to modify Algorithm 2 by allowing the shapes of the minimal and optimal penalties to be different, hence the following algorithm.

Algorithm 3.
`Input`: $\big(P_n\gamma(\widehat{s}_m)\big)_{m\in\mathcal{M}}$, $\big(\mathrm{pen}_0(m)\big)_{m\in\mathcal{M}}$, $\big(\mathrm{pen}_1(m)\big)_{m\in\mathcal{M}}$ and $(\mathbb{C}_m)_{m\in\mathcal{M}}$.

(1) Compute $\big(\widehat{m}_{\min}(C)\big)_{C\geq 0}$, where for every $C \geq 0$,

$$\widehat{m}_{\min}(C) \in \operatorname*{argmin}_{m\in\mathcal{M}}\big\{P_n\gamma(\widehat{s}_m) + C\,\mathrm{pen}_0(m)\big\} \ . \tag{4.32}$$

(2) Find $\widehat{C}_{\mathrm{jump}} > 0$ corresponding to the "unique large jump" of $C \mapsto \mathbb{C}_{\widehat{m}_{\min}(C)}$.

(3) Select

$$\widehat{m}_{\mathrm{Alg.3}} \in \operatorname*{argmin}_{m\in\mathcal{M}}\Big\{P_n\gamma(\widehat{s}_m) + \widehat{C}_{\mathrm{jump}}\,\mathrm{pen}_1(m)\Big\} .$$

`Output`: $\widehat{m}_{\mathrm{Alg.3}}$.

Algorithm 3 implicitly assumes that the minimal and the optimal penalty are respectively equal to $C^\star\,\mathrm{pen}_0$ and $C^\star\,\mathrm{pen}_1$, with pen_0 and pen_1 known but $C^\star$ unknown.

For linear estimators, the above arguments suggest to apply Algorithm 3 with

$$\widehat{s}_m = \widehat{F}_m \qquad P_n\gamma(\widehat{s}_m) = \frac{1}{n}\Big\|\widehat{F}_m - Y\Big\|^2 \qquad \mathbb{C}_m = \mathrm{tr}(A_m)$$
$$\mathrm{pen}_0(m) = \frac{2\,\mathrm{tr}(A_m) - \mathrm{tr}\big(A_m^\top A_m\big)}{n} \qquad \text{and} \qquad \mathrm{pen}_1(m) = \frac{2\,\mathrm{tr}(A_m)}{n} \ . \tag{4.33}$$

This procedure works well, as illustrated by Figure 3 as well as the theoretical results and the numerical experiments of **[11]** and **[18]**. In particular, for Algorithm 3 with linear estimators, a result similar to Theorem 4.1 is proved in **[18]** with the following few changes:

- Some assumptions on the matrices A_m must be added: for all $m \in \mathcal{M}$,

$$\mathrm{tr}\big(A_m^\top A_m\big) \leq \mathrm{tr}(A_m)\,, \qquad \mathrm{tr}(A_m) \leq n \qquad \text{and} \qquad |\!|\!|A_m|\!|\!| \leq M$$

  some constant, where $|\!|\!|\cdot|\!|\!|$ denotes the operator norm.
- Assumption (**HO**) is replaced by a stronger one: some $m_2 \in \mathcal{M}$ exists such that

$$\mathrm{tr}(A_{m_2}) \leq \sqrt{n} \qquad \text{and} \qquad \big\|(I_n - A_m)F\big\|^2 \leq \sigma^2\sqrt{n\log(n)}\,.$$

- In Eq. (4.12)–(4.13), the dimension D_m is replaced by the degrees of freedom $\mathrm{tr}(A_m)$, and the lower bound in Eq. (4.12) is $n/3$ instead of $9n/10$.
- The deviation bounds η_n^-, η_n^+ are of order $\sqrt{\log(n)/n}$, and the remainder terms in Eq. (4.14) are slightly enlarged.

Note also that the results of **[18]** allow to consider a continuous family $\mathcal{M}$, for choosing a regularization parameter in kernel ridge regression with a fixed kernel.

Comparing the shapes of the minimal and optimal penalties in Eq. (4.33) enlightens the slope heuristics (4.11): the factor 2 between the minimal and optimal penalty corresponds to the relationship $\mathrm{pen}_1 = 2\,\mathrm{pen}_0$, which holds for linear estimators if and only if for all $m \in \mathcal{M}$,

$$\mathrm{tr}\big(A_m^\top A_m\big) = \mathrm{tr}(A_m) \ .$$

For projection estimators, this always holds true because $A_m^\top A_m = A_m$. Suprisingly, one always has $\mathrm{tr}(A_m^\top A_m) = \mathrm{tr}(A_m)$ for k-nearest neighbors also **[18]**. In other cases, such as kernel ridge or Nadaraya-Watson estimators, the minimal and optimal penalties are never proportional, except maybe in some very specific cases, and we only have $\mathrm{pen}_1(m)/\,\mathrm{pen}_0(m) \in (1, 2]$.

Cross-validation methods could also be used for selection among linear estimators in regression. They would certainly satisfy a similar oracle inequality (at first order at least), even if no result of this kind has been proved up to now. Nevertheless, experiments in **[18]** show that cross-validation performs equally well or worse than the above penalization methods, because penalization here takes advantage of some specificities of the problem, and of the knowledge that the noise-level is constant. Since cross-validation requires a much larger computational cost, choosing among the two procedures is straightforward in that setting.

4.7. Minimal penalties in other settings

The slope heuristics and Algorithm 2 are theoretically validated in a few more frameworks, mostly for least-squares estimators (in regression or in density estimation) and "close to least-squares" estimators (for instance, histogram density estimation with the maximum-likelihood contrast [**Sau10b**] or, more generally, "regular" estimators [**Sau10a**]), see **[22]**. In particular, the slope heuristics can work outside the regression setting [**Ler12**] or with dependent data [**Ler11**].

In **[1]**, the slope heuristics is validated for random-design regression and with heteroscedastic data—that is, when $\mathbb{E}[\varepsilon^2 \,|\, X] = \sigma^2(X)$ can vary with X—, for regressogram estimators. Although the conclusion is similar to that of Theorem 4.1, the proof of the result of **[1]** is technically much more complex. In particular, computing expectations of the key quantities as in Eq. (4.5)–(4.6) is not at all straightforward, first because of heteroscedasticity, and also because of specificities of the random-design framework. Note finally that for heteroscedastic regression, the shape of the optimal and minimal penalties depend on the function $\sigma^2(\cdot)$, which is unknown in general, so it must be estimated by resampling **[3]**. The slope heuristics is useful in that context because resampling-based penalties involve multiplicative constants whose non-asymptotic optimal value is not theoretically known in general, as argued in the third point in Section 4.1.

Up to now, linear estimators are the only known example for which Algorithm 3 must be used instead of Algorithm 2, but we believe such a generalization will be useful in other settings, see Section 7.3.

More generally, the slope heuristics and its generalization to minimal penalties (Algorithm 3) suggest that phase transitions can be useful for building estimator selection procedures, if such clearly observable phenomena can be related to the key unknown parameters of the problem.

CHAPTER 5

Change-point detection

Change-point detection, also called one-dimensional segmentation, is a classical statistical problem: given a time series $Y_1, \dots, Y_n \in \mathcal{Y}$ whose distribution abruptly changes at some unknown instants, the goal is to recover the number of these changes and their locations. This problem is motivated by a wide range of applications, such as audio processing [**HVLYFC09**], financial time-series analysis [**LT06**] and Comparative Genomic Hybridization (CGH) data analysis [**Pic05**]. A large literature exists about change-point detection in many frameworks, see [**BD93**, **TNM14**] for a bibliography.

The most classical examples are when $\mathcal{Y} = \mathbb{R}$, the Y_i are assumed independent, and the goal is to recover (i) change-points in the mean $\mathbb{E}[Y_i]$, assuming the variance of the Y_i is constant, or (ii) change-points in both the mean and the variance. The papers **[6]**, **[19]** and **[12]** address less well understood situations that correspond to many applications: detecting change-points in the mean $\mathbb{E}[Y_i]$ without assuming the variance of the Y_i is constant **[6]**, detecting change-points in the full distribution of $Y_i \in \mathbb{R}$ that do not imply any change of the mean or the variance **[19]**, and the case where the Y_i are high-dimensional **[12]** or non-vectorial **[19]**. In all three papers, the

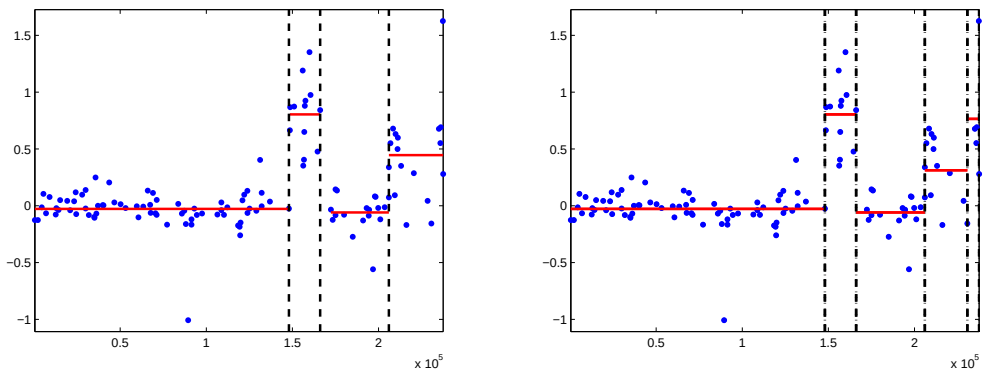


FIGURE 1. The change-point detection problem for CGH data [**Pic05**]: given a time series (blue points), find its change-points. Vertical dashed-lines show the estimated breakpoints, red lines show the corresponding estimates of the mean on each segment. Left: procedure of **[6]**. Right: penalization procedure (5.3).

change-point detection problem is cast into the model selection framework, as done for instance in [**Yao88**, **YA89**, **CR04**, **Lav05**, **Leb05**, **BKL+09**].

5.1. Change-point detection and model selection

Assume $Y_1, \dots, Y_n \in \mathcal{Y} = \mathbb{R}$ are independent, and the goal is to recover changes in the mean of the Y_i. As in the fixed-design regression framework (Section 4.2), we can write

$$Y_i = F_i + \varepsilon_i$$

where $\varepsilon_1, \dots, \varepsilon_n$ are independent with zero-mean and $F \in \mathbb{R}^n$ is unknown.

Recovering the change-points of F is equivalent to recovering the smallest segmentation of $\{1, \dots, n\}$ such that F_i is constant on each of its segments. More precisely, let $\mathcal{M}_n$ be the set of segmentations of $\{1, \dots, n\}$, that is the set of partitions of $\{1, \dots, n\}$ of the form

$$\big\{\{1, \dots, k_1\}, \{k_1 + 1, \dots, k_2\}, \dots, \{k_{D-1} + 1, \dots, n\}\big\} \tag{5.1}$$

with $D \geq 1$ and $1 \leq k_1 < \cdots < k_{D-1} \leq n$. For every $m \in \mathcal{M}_n$, let S_m denote the space of functions $\{1, \dots, n\} \to \mathbb{R}$ that are constant on each segment of m. Seeing S_m as a subspace of $\mathbb{R}^n$ and denoting by $(e_1, \dots, e_n)$ the canonical basis of $\mathbb{R}^n$, if m is defined by Eq. (5.1), then S_m is generated by

$$(e_1 + \cdots + e_{k_1}, e_{k_1+1} + \cdots + e_{k_2}, \dots, e_{k_{D-1}+1} + \cdots + e_n) \ .$$

Then, recovering the change-points of F is equivalent to finding the smallest segmentation $m^\star \in \mathcal{M}_n$ such that $F \in S_{m^\star}$, which is a model selection problem (with an identification goal). Even if estimation and identification are different goals for model selection, a classical approach is to use for change-point detection a model selection procedure for which an oracle inequality (2.2) can be proved with the least-squares loss of Section 4.2

$$\mathcal{L}\left(\widehat{F}_m\right) = \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 \quad \text{where} \quad \widehat{F}_m = \Pi_{S_m} = \Pi_m Y \ .$$

The underlying idea is that the least-squares loss quantifies well the importance of having detected each change-point, since the signal-to-noise ratio is often too small in practice for hoping to detect precisely all change-points.

The collection $(S_m)_{m \in \mathcal{M}_n}$ is "large" according to the terminology of Section 2.2, since $\mathrm{Card}(\mathcal{M}_n) = 2^{n-1}$. Therefore, following Section 2.2.1, one can build a change-point detection procedure by penalization

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}_n} \left\{ \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + \mathrm{pen}(m) \right\} \tag{5.2}$$

by choosing $\mathrm{pen} : \mathcal{M}_n \to \mathbb{R}$ such that

$$\forall m \in \mathcal{M}_n, \quad \mathrm{pen}(m) \geq \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 - \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2$$

holds with a large probability, as done in [**BM01**] for general model selection problems and in [**CR04**, **Lav05**, **Leb05**] for change-point detection. In particular, assuming that

$$\mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2$$

does not depend on i, based on a combination of theoretical and experimental arguments, [**Leb05**] suggests the penalty

$$\mathrm{pen}(m) = \frac{5\sigma^2 D_m}{n} + \frac{2\sigma^2 D_m}{n} \log\left(\frac{n}{D_m}\right) , \tag{5.3}$$

where $D_m = \mathrm{Card}(m)$ is the dimension of S_m. Note that such a penalization procedure can be computed efficiently thanks to dynamic programming, and pruning techniques can fasten the algorithm [**Rig10**].

5.2. Heteroscedastic data

Penalties of the form of Eq. (5.3) have been proved to work well for change-point detection in several papers. Nevertheless, they clearly assume that the variance of the data is constant along the time series, an assumption that is often violated in practice, for instance with CGH data [**Pic05**]. Furthermore, **[17]** shows that dimensionality-based penalties fail in heteroscedastic regression, for model selection among a "polynomial" family, compared to some resampling-based methods **[3]** for instance. Therefore, if one wants to detect only change-points in the mean with no constraint on the variance, another approach has to be proposed.

In **[6]**, a cross-validation based change-point detection procedure is shown to solve this problem satisfactorily on some synthetic data, with an application to the CGH data of [**Pic05**], see Figure 1. The starting point is the remark that using a penalty such as the one of Eq. (5.3), which depends on m only through D_m, leads to a procedure that can be reformulated as the following two-step procedure:

(1) For every D, choose

$$\widehat{m}(D) \in \operatorname*{argmin}_{m \in \mathcal{M}_n(D)} \left\{ \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 \right\} \tag{5.4}$$

where $\mathcal{M}_n(D) := \{m \in \mathcal{M}_n \,/\, D_m = D\}$.

(2) Then, choose $\widehat{m}\big(\widehat{D}\big)$ where

$$\widehat{D} \in \operatorname*{argmin}_{D \in \{1,\dots,n\}} \left\{ \frac{1}{n} \left\| \widehat{F}_{\widehat{m}(D)} - Y \right\|^2 + \mathrm{pen}(D) \right\} .$$

When the data are heteroscedastic, comparing the expectations of the empirical risk $n^{-1}\|\widehat{F}_m - Y\|^2$ and of the loss $n^{-1}\|\widehat{F}_m - F\|^2$ shows that (5.4) should lead to assigning preferentially change-points to areas where the noise level is large, that is, to overfit,

see Lemma 1 in [6]. Numerical experiments confirm this failure, and that it cannot be compensated in the second step of the procedure when selecting $\widehat{D}$.

The computation of the expectation of cross-validation (CV) criteria with regressogram estimators in [16] suggests that replacing the empirical risk in Eq. (5.4) by a CV criterion should lead to a much better segmentation $\widehat{m}(D)$, assuming the number of change-points is known. A key issue is to be able to perform efficiently the minimization of the CV estimator of the risk of $\widehat{F}_m$ over $\mathcal{M}_n(D)$, which is huge if $D \geq 3$ unless n is very small. This can be done roughly with the same complexity as Eq. (5.4) thanks to dynamic programming—which can still be used here—and closed-form formulas for leave-p-out estimators of the risk of regressograms [Cel08]. The remaining issue is to select D, that is, to choose among the family $(\widehat{s}_{\widehat{m}(D)})_{1\leq D\leq n}$ of estimators. Since this family is polynomial—it contains n estimators—, the unbiased risk estimation principle can be applied (see Section 2.2.2), for instance using V-fold CV. Overall, the new change-point detection procedure proposed in [6] is the following:

(1) For every D, choose

$$\widehat{m}_1(D) \in \operatorname*{argmin}_{m\in\mathcal{M}_n(D)} \left\{ \widehat{\mathcal{L}}^{\mathrm{CV}}\left(\widehat{F}_m; (i, Y_i)_{1\leq i\leq n}\right)\right\} ,$$

for some CV criterion $\widehat{\mathcal{L}}^{\mathrm{CV}}$, see Section 3.1.

(2) Then, choose $\widehat{m}_1\big(\widehat{D}_2\big)$ where

$$\widehat{D}_2 \in \operatorname*{argmin}_{D\in\{1,\dots,n\}} \left\{ \widehat{\mathcal{L}}^{\mathrm{VF}}\left(\widehat{F}_m; (i, Y_i)_{1\leq i\leq n}\right)\right\}$$

for some V-fold CV criterion, see Section 3.1.

Numerical experiments in [6] show that this procedure improves significantly over existing methods when the data are heteroscedastic, taking for instance leave-one-out at step 1 and 5-fold CV at step 2.

5.3. High-dimensional or complex data

The approach of Sections 5.1–5.2 can be straightforwardly extended to multivariate data $Y_i \in \mathcal{Y} = \mathbb{R}^d$ for any $d \geq 1$, when the goal is to detect changes in the mean. Other procedures have also been proposed for change-point detection in a high-dimensional time series [PLBR11, BV11]. Nevertheless, in many application domains, data have no vectorial structure, but are represented as histograms (for instance, in audio processing or computer vision), strings (for instance, texts or DNA sequences) or graphs (for instance, in social science or bioinformatics). Even when $\mathcal{Y} = \mathbb{R}^d$ with d large, using implicitly the Euclidean metric for measuring the distance between two observations—as done in Sections 5.1–5.2—is not appropriate since it does not take into account the structure of data. Finally, when $\mathcal{Y} = \mathbb{R}$, change-points can occur in other features of the

data than mean and variance. In all three cases, the approach of Sections 5.1–5.2 cannot be used, and **[19]** proposes a general methodology for these change-point detection problems.

In a few words, the idea of **[19]** is to transform the original data $Y_1, \dots, Y_n \in \mathcal{Y}$ into

$$\Phi(Y_1), \dots, \Phi(Y_n) \in \mathcal{H}$$

some Hilbert space, and then to apply the approach of Section 5.1 to this transformed time series, which is possible since $\mathcal{H}$ has an Hilbertian structure. More precisely, given some positive semi-definite kernel $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, we define $\mathcal{H}$ its associated reproducing kernel Hilbert space and $\Phi : y \in \mathcal{Y} \mapsto k(y, \cdot) \in \mathcal{H}$ the canonical feature map [**SS01**, **SC08**]. Assuming for simplicity that the kernel k is bounded, one can write

$$\Phi(Y_i) = Z_i = F_i + \varepsilon_i \qquad \text{with} \qquad \forall f \in \mathcal{H}, \quad \mathbb{E}\big[\langle \varepsilon_i, f \rangle_{\mathcal{H}}\big] = 0$$

as in Section 5.1; $F_i \in \mathcal{H}$ is called the mean-element of the distribution of Y_i. Up to technical issues related to the fact that Z_i, F_i and ε_i belong to $\mathcal{H}$ instead of $\mathbb{R}$, the arguments of Section 5.1 can be extended, leading to a change-point detection procedure of the form

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|_{\mathcal{H}^n}^2 + \frac{C v_{\max} D_m}{n} \left[1 + \log\left(\frac{n}{D_m} \right) \right] \right\} , \tag{5.5}$$

where $v_{\max}$ is an upper bound on the "variance" of the Z_i. Obviously, the above summary skips some key issues, we refer to **[19]** for a detailed presentation of the procedure.

A key remark is that $\widehat{m}$ defined by Eq. (5.5) can be computed efficiently thanks to the combination of dynamic programming—as in Section 5.1—and the kernel trick, which allows to avoid computations in $\mathcal{H}$ since the penalization procedure (5.3) only requires to compute dot products between observations. The latter point is important since $\mathcal{H}$ can be infinite-dimensional.

Then, **[19]** proves an oracle inequality for (5.5), which guarantees it estimates well F in $\mathcal{H}$; we conjecture that this implies good change-point detection performance provided the kernel k is adapted to the problem considered. Compared to similar results proved in the real case ($\mathcal{Y} = \mathbb{R}$), three main additional difficulties arise. First, the Z_i belong to some Hilbert space, in which computations of expectations and concentration results are more difficult to obtain than in $\mathbb{R}$. Second, the usual Gaussian assumption on the noise is meaningless, so it is replaced by a boundedness assumption, which requires to prove new concentration inequalities. Third, we can no longer assume that data are homoscedastic, in particular because for translation invariant kernels, the "variance" of the Z_i is directly related to the norm of the mean elements F_i in $\mathcal{H}$, which has no reason to stay constant when F_i jumps. This explains why σ^2 in the penalty (5.3) has been changed into $v_{\max}$ in Eq. (5.5).

Experiments in **[19]** on synthetic and on real data show the advantage of this procedure for the three problems mentioned at the beginning of the section. In particular,

using the Gaussian kernel with real data, one can detect changes in distributions even when both the mean and the variance are constant; an application to real data—audio and video streams—is also presented in **[19]**.

5.4. Metric learning for multivariate data

When $\mathcal{Y} = \mathbb{R}^d$, the classical empirical risk minimization approach (5.2) can be naturally generalized, with the multivariate least-squares contrast $\gamma\big(t;(x,y)\big) = \big\|t(x)-y\big\|^2_{\mathbb{R}^d}$. Nevertheless, using the Euclidean metric on $\mathbb{R}^d$ might not be appropriate, especially when d is large. For instance, some of the d coordinates can be non-informative about the target change-points, or even worse, some coordinates can suggest to place change-points at locations where the phenomenon of interest does not change. In such cases, we would like to be able to simply discard these non-informative or perturbing coordinates.

If we knew an appropriate (pseudo)metric on $\mathbb{R}^d$ for the change-point detection problem we want to solve, given by some matrix $B \in \mathcal{S}_d^+$ the set of $d \times d$ symmetric positive semi-definite matrices, we would replace Eq. (5.2)–(5.3) by

$$\widehat{m}_B(Y) \in \operatorname*{argmin}_{m\in\mathcal{M}_n}\left\{\frac{1}{n}\sum_{i=1}^{n}\Big\|\big(\widehat{F}_m\big)_i - Y_i\Big\|_B^2 + D_m\right\} \tag{5.6}$$

$$\text{where} \qquad \forall u \in \mathbb{R}^d,\ \|u\|_B^2 = u^\top B u \ .$$

Note that in Eq. (5.6), we simplify the penalty (5.3) by taking instead a penalty proportional to the dimension λD_m for some $\lambda > 0$; the parameter λ can be set to 1 because multiplying the penalty D_m by $\lambda > 0$ is equivalent to dividing the matrix B by λ, hence definition (5.6). The minimization problem in Eq. (5.6) can be solved efficiently thanks to dynamic programming, see **[12]**.

In general, B is unknown and must be learned. In the "metric learning" setting, which is considered in **[12]**, we assume some supervised data are available, that is, j time series $Y^1, \ldots, Y^j$—of respective lengths $n_1, \ldots, n_j$, not necessarily equal—, together with the corresponding segmentations $m^1, \ldots, m^j$. We assume that these j time series are "similar" to the new ones we want to segment, that is, they are all well segmented with the same (unknown) metric B. So, we can use these labeled data $(Y^i, m^i)_{1\le i\le j}$ for learning B. From a practical point of view, having such supervised data seems much more realistic than assuming B is known: we can always ask an expert of a given field to describe on some examples where are the breakpoints we are looking for, but not to provide us directly a good metric B.

Then, given some contrast [1] function γ over segmentations—see **[12]** for examples—the classical regularized empirical risk approach suggests to minimize over $B \in \mathcal{S}_d^+$ the

1. In Section 2.1.2, a contrast function is defined as a function $\mathbb{S} \times \Xi \to \mathbb{R}$, but in the prediction framework the contrast at $(t;(x,y))$ can usually be written as $\gamma(t(x);y)$ for some function $\gamma : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$. Here, by abuse of terminology, we use the term contrast for such a function $\mathcal{M}_n \times \mathcal{M}_n \to \mathbb{R}$.

criterion

$$\frac{1}{j}\sum_{i=1}^{j}\gamma\Big(m^i, \widehat{m}_B\big(Y^i\big)\Big) + \Omega(B) \ , \tag{5.7}$$

where Ω is some regularization term. A classical choice for Ω is the Frobenius norm $\Omega(B) = \mathrm{tr}\big(B^\top B\big)$, but one can also take the trace instead, in order to obtain a low-rank matrix B.

Minimizing exactly the criterion given by Eq. (5.7) is untractable, so **[12]** suggests a convex relaxation of it that can be minimized efficiently. This relaxation takes advantage of a particular formulation of the problem, which allows to use the large-margin approach that was proposed first by [**THJA05**] for structured prediction problems. Structured prediction corresponds to the prediction problem (Example 2.1) when the variable of interest is a combinatorial structured object. In the setting of **[12]**, the variables of interest are the segmentations $m^1, \ldots, m^j$ and the goal is to learn some metric B such that $\widehat{m}_B\big(Y^i\big)$ "predicts" well the segmentation m^i.

Then, experiments on synthetic and on real data in **[12]** show the good performance of the method, even when only partial information about the segmentations $m^1, \ldots, m^j$ is available.

CHAPTER 6

Estimator selection for some other learning problems

This chapter describes some other works on estimator selection procedures (or related issues) for various statistical learning problems: unsupervised learning problems in [12] and [13], regression with random forests in [20], multivariate (or multi-task) regression in [9], margin adaptive model selection in binary classification in [8], and multi-label classification in [21].

6.1. Metric learning for unsupervised learning

The metric learning problem for change-point detection is presented in Section 5.4. It can be extended to several unsupervised learning problems, as done in [12] and [13] for instance.

Unsupervised learning with multivariate observations can be described as follows. Let $d \geq 1$ be fixed and assume that we observe some $Y = (Y_1, \ldots, Y_n) \in (\mathbb{R}^d)^n$. The goal is to infer from Y some parameter $m \in \mathcal{M}$ of its generating process. For instance, in change-point detection (see Chapter 5), $m^\star$ is a segmentation of $\{1, \ldots, n\}$; in clustering, $m^\star$ is a partition of $\{1, \ldots, n\}$ such that all clusters $(Y_i)_{i \in \lambda}$, $\lambda \in m^\star$ are homogeneous.

Any B in $\mathcal{S}_d^+$ the set of $d \times d$ symmetric positive semi-definite matrices defines a (pseudo)metric on $\mathbb{R}^d$. Assume that for every $B \in \mathcal{S}_d^+$, some unsupervised learning procedure $\widehat{m}_B$ is available, that estimates m with $\widehat{m}_B(Y)$. Then, learning the metric B is an estimator selection problem, among the family $(\widehat{m}_B)_{B \in \mathcal{S}_d^+}$.

Metric learning tackles this problem by assuming some supervised data are available. More precisely, we assume that independent instances $(Y^i, m^i)_{1 \leq i \leq j}$ are given such that for every $i \in \{1, \ldots, j\}$, m^i is the parameter of interest associated with Y^i. The problem is to choose some metric $\widehat{B}$ such that for any "new" data (Y^{j+1}, m^{j+1}), given Y^{j+1} only, the associated m^{j+1} is well estimated by $\widehat{m}_{\widehat{B}}(Y^{j+1})$. Assuming that a contrast function $\gamma : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ is given, the goal is to minimize (in expectation)

$$\gamma\Big(m^{j+1}, \widehat{m}_B\big(Y^{j+1}\big)\Big)$$

over $B \in \mathcal{S}_d^+$, which is a prediction problem (Example 2.1) with $\mathcal{X} = (\mathbb{R}^d)^n$ and $\mathcal{Y} = \mathcal{M}$.

As in Section 5.4, a natural procedure would be

$$\widehat{B} \in \operatorname*{argmin}_{B \in \mathcal{S}_d^+} \left\{ \frac{1}{j} \sum_{i=1}^{j} \gamma\Big(m^i, \widehat{m}_B\big(Y^i\big)\Big) + \Omega(B) \right\} \tag{6.1}$$

for some regularization term $\Omega : \mathcal{S}_d^+ \to \mathbb{R}^+$, but the optimization problem (6.1) is usually untractable and the challenge is to find some approximate solution to it.

In addition to the change-point detection case described in Section 5.4, **[12]** tackles this issue for other partitioning problems: two-dimensional segmentation (with an application to image segmentation) and clustering. In **[13]**, the same problem is addressed for dynamic time warping, which is the problem of aligning two (multivariate) time series, that arises for instance in audio processing [**CSS+07**] and in bioinformatics [**TPP99**].

In both cases, the large-margin approach of [**THJA05**] for structured prediction is used for proposing a computationally tractable relaxation of problem (6.1). Experiments on synthetic and on real data show the good performance of the relaxation, compared to the state of the art.

6.2. Approximation error rates of purely random forests

Random forests [**Bre01**] are popular for classification and regression problems. Let us describe briefly how they are defined in regression, using the notation of Example 2.2 and assuming $\mathcal{X} \subset \mathbb{R}^d$. A more detailed presentation and some references can be found in **[20]** and [**Gen10**]. Let D_n be some sample. For every partition m of $\mathcal{X}$, let $\widehat{s}_m(D_n)$ be the corresponding regressogram estimator: on each element λ of the partition m, the function $\widehat{s}_m(D_n)$ is constant, equal to the average of the Y_i such that $(X_i, Y_i) \in D_n$ and $X_i \in \lambda$. Given D_n and some random process $\theta \sim \Theta$ independent from D_n, some random partition $m(\theta; D_n)$ is defined, usually associated to a decision tree. The resulting estimator $\widehat{s}_{m(\theta;D_n)}(D_n)$ is then called a (random) tree. Given D_n and a sequence $\theta_1, \dots, \theta_q$ of i.i.d. random variables with distribution Θ, independent from D_n, the associated *random forest* estimator is defined as the average of the corresponding trees:

$$\widehat{s}\,(\theta_{1\dots q}; D_n) = \frac{1}{q} \sum_{j=1}^{q} \widehat{s}_{m(\theta_j;D_n)}(D_n) \ .$$

Random forests perform remarkably well in practice, but showing theoretical results on the model proposed by [**Bre01**] remains a widely open problem, see [**SBV14**] for some recent results. As a first step, purely random forests [**Bre00**] have been introduced, where the random partitions $m(\theta; D_n)$ are assumed to be independent from D_n. This assumption simplifies much the theoretical analysis—which remains challenging, in particular because the $\big(\widehat{s}_{m(\theta_j)}(D_n)\big)_{1\le j\le q}$ are dependent—, and a few theoretical guarantees have been obtained on such forests, see [**BDL08**, **Bia12**, **Gen12**].

In particular, the least-squares loss of purely random forests can be decomposed into an approximation and an estimation error, as in Eq. (2.3). For every partition m of $\mathcal{X}$, we define

$$s_m^\star = \sum_{\lambda \in m} \beta_\lambda \mathbb{1}_\lambda \quad \text{where} \quad \beta_\lambda := \mathbb{E}\big[s^\star(X) \,|\, X \in \lambda\big] \ ,$$

and for every sequence $\theta_1, \dots, \theta_q$ (defining a forest),

$$s^\star_{\theta_{1\dots q}} = \frac{1}{q}\sum_{j=1}^{q} s^\star_{m(\theta_j)} \ .$$

Proposition 1 in **[20]** shows that for every $x \in \mathcal{X}$

$$\begin{aligned} &\mathbb{E}\Big[\big(s^\star(x) - \widehat{s}\,(x;\theta_{1\dots q};D_n)\big)^2\Big] \\ &= \underbrace{\mathcal{B}_{\Theta,\infty}(x) + \frac{\mathcal{V}_\Theta(x)}{q}}_{\text{approximation error or bias}} + \underbrace{\mathbb{E}\left[\left(s^\star_{\theta_{1\dots q}}(x) - \widehat{s}\,(x;\theta_{1\dots q};D_n)\right)^2\right]}_{\text{estimation error or variance}} \end{aligned} \tag{6.2}$$

$$\text{with} \quad \mathcal{B}_{\Theta,\infty}(x) := \Big(s^\star(x) - \mathbb{E}_{\theta\sim\Theta}\Big[s^\star_{m(\theta)}(x)\Big]\Big)^2 \quad \text{and} \quad \mathcal{V}_\Theta(x) := \mathrm{var}_{\theta\sim\Theta}\Big(s^\star_{m(\theta)}(x)\Big) \ .$$

Eq. (6.2) can also be integrated with respect to $x = X$ in order to get a result fully comparable with Eq. (2.3).

Jensen's inequality shows that the estimation error in Eq. (6.2) is smaller for a forest than for a single tree. For a particular partitioning process Θ, **[Gen12]** showed that the estimation error is strictly smaller for an infinite forest than for a single tree, by a multiplicative factor 3/4. In particular, this shows that considering a forest can strictly improve the performance compared to a single tree.

What about the approximation error term in Eq. (6.2)? Since a variance is non-negative, it decreases with the number q of trees in the forest, strictly if $s^\star_{m(\theta)}(x)$ is not deterministic. Can we expect a significant improvement, for instance in the approximation *rate* as a function of the size of the partitions? The answer is positive, as proved in **[20]** for three partitioning processes Θ, which implies in particular that forest estimators with enough trees attain a strictly better learning rate than single tree estimators.

More precisely, assuming that the regression function $s^\star$ is smooth enough—twice or three times differentiable—and that X follows a uniform distribution over $\mathcal{X} = [0,1]^d$, **[20]** proves general upper and lower bounds for $\mathcal{B}_{\Theta,\infty}(x)$ and $\mathcal{V}_\Theta(x)$, which depend on the derivatives of $s^\star$ and on the properties of Θ. The key quantities it involves are the moments of $x_i - A_{i,\theta}(x)$ and $B_{i,\theta}(x) - x_i$, for $i = 1, \dots, d$, where $\prod_{i=1}^d \big[A_{i,\theta}(x), B_{i,\theta}(x)\big)$ is the unique element of $m(\theta)$ to which x belongs.

Then, for three partitioning processes Θ in which $\mathrm{Card}(m(\theta)) = k$ is deterministic, **[20]** shows that

$$\mathcal{B}_{\Theta,\infty}(x) \lesssim k^{-2\alpha} \quad \text{and} \quad \mathcal{V}_\Theta(x) \gtrsim k^{-\alpha} \tag{6.3}$$

for some $\alpha > 0$, except maybe for x too close to the boundaries of $\mathcal{X}$. As a consequence, assuming that the size k of the trees can be chosen optimally—which is an estimator selection problem that can be solved in practice by using cross-validation for instance,

see Chapter 3—, the risk of a single tree estimator is lower bounded as

$$\mathbb{E}\Big[\big(s^\star(x) - \widehat{s}\,(x;\theta_1;D_n)\big)^2\Big] \gtrsim \inf_{1\le k\le n}\left\{k^{-\alpha} + \frac{\sigma^2 k}{n}\right\} \gtrsim \left(\frac{\sigma^2}{n}\right)^{-\frac{\alpha}{\alpha+1}} \tag{6.4}$$

while the risk of a large enough forest is upper bounded as

$$\mathbb{E}\Big[\big(s^\star(x) - \widehat{s}\,(x;\theta_{1\dots q};D_n)\big)^2\Big] \lesssim \inf_{1\le k\le n}\left\{k^{-2\alpha} + \frac{\sigma^2 k}{n}\right\} \lesssim \left(\frac{\sigma^2}{n}\right)^{-\frac{2\alpha}{2\alpha+1}} . \tag{6.5}$$

In particular, comparing Eq. (6.4) and (6.5) shows that a forest attains a better learning rate than a tree. For the first two processes Θ considered in **[20]**, which assume $d = 1$, Eq. (6.3)–(6.5) hold with $\alpha = 2$ if $s^\star$ is twice differentiable, so that a large enough forest attains the corresponding minimax learning rate [**GKKW02**]. For the third process Θ considered in **[20]**, which is defined for any $d \ge 1$, Eq. (6.3)–(6.5) hold with

$$\alpha = \frac{-\log\big(1 - \frac{1}{2d}\big)}{\log(2)} < \frac{2}{d}$$

if $s^\star$ is twice differentiable, so even an infinite forest does not attain the minimax learning rate $n^{-4/(d+4)}$. Its learning rate is still significantly better than the one of a single tree.

Experiments on synthetic data in **[20]** confirm the above theoretical statements and study the approximation error rates of a fourth purely random forest algorithm from [**Bia12**], called "Hold-out random forests" in **[20]**. Hold-out random forests are interesting because they are quite close to the original random forest model of Breiman [**Bre01**]: the random partitions used for building each tree are $m(\theta; D'_n)$ for the same process as Breiman's random forest, except that it is applied to an additional sample D'_n independent from D_n. The conclusion of the experiments is that hold-out random forests seem to attain better approximation rates than single trees, by a factor ≈ 1.6 in the examples considered in **[20]**. We can thus conjecture that the good practical performance of random forests is—at least partly—due to an improvement of the approximation properties induced by the averaging over a large family of regressogram estimators.

Note finally that the above theoretical results provide an answer to some practical problem: how many trees q should be put in a forest? According to Eq. (6.2) and (6.3), the learning rate of Eq. (6.5) is attained as soon as $q \ge (k^\star)^\alpha$ where $k^\star \propto (n/\sigma^2)^{1/(2\alpha+1)}$.

6.3. Multi-task kernel ridge regression

Multi-task learning is motivated by the fact that increasing the sample size is often difficult in practice—no more labelled data can be found, or they are costly to obtain—, but some data are available for similar learning problems. For instance, image classification for 'dogs' and for 'cats' certainly share some similarities. The multi-task paradigm assumes one can increase the "effective" sample size by solving similar tasks together.

Nevertheless, analyzing theoretically such multi-task techniques requires to formalize clearly what are "similar" tasks.

Multi-task (or multivariate) regression, that is, Example 2.2 with $\mathcal{Y} = \mathbb{R}^p$ for some $p \geq 1$, is tackled in [9]. Estimating each coordinate of the variable of interest Y is a regression task, and performing the estimation jointly over the p tasks should provide better performance than considering them separately, provided we can take advantage from their similarity. As in Example 2.2, we can write that for every $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, p\}$,

$$Y_i^j = \eta^j(X_i) + \varepsilon_i^j \tag{6.6}$$

where $\eta^1, \dots, \eta^p : \mathcal{X} \to \mathbb{R}$ are the p regression functions, and the vectors

$$\varepsilon_i = \left(\varepsilon_i^j\right)_{1 \leq j \leq p} \in \mathbb{R}^p,\ i = 1, \dots, n\,,$$

are independent with zero mean and (unknown) covariance matrix Σ. We consider the least-squares loss, so that the goal is to estimate $\eta^1, \dots, \eta^p$.

One "multi-task assumption" that can be considered in [9]—among others—is that the functions $\eta^1, \dots, \eta^p$ are close in some functional space $\mathcal{H}$. For such a multi-task problem, [**EMP05**] proposed the following multi-task kernel ridge estimator, assuming that $\mathcal{H}$ is a reproducing kernel Hilbert space with positive-definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$: for every $\lambda, \mu \in (0, +\infty)^2$,

$$\widehat{s}_{\lambda,\mu} \in \operatorname*{argmin}_{g \in \mathcal{H}^p} \left\{ \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(Y_i^j - g^j(X_i)\right)^2 + \lambda \sum_{j=1}^{p} \left\|g^j\right\|_{\mathcal{H}}^2 + \frac{\mu}{2} \sum_{j=1}^{p} \sum_{k=1}^{p} \left\|g^j - g^k\right\|_{\mathcal{H}}^2 \right\}. \tag{6.7}$$

In Eq. (6.7), the least-squares empirical risk is regularized by two terms. The sum of norms $\|g^j\|_{\mathcal{H}}^2$ is the usual regularization term in (single-task) kernel ridge regression; it enforces the functions g^j to be smooth, hence to avoid overfitting, provided λ is large enough. The sum of norms $\|g^j - g^k\|_{\mathcal{H}}^2$ enforces the functions g^j to be close to each other in $\mathcal{H}$, which is exactly our multi-task assumption on $\eta^1, \dots, \eta^p$.

A generalization of Eq. (6.7) is the following: given some matrix $M \in \mathcal{S}_p^{++}$ the set of $p \times p$ symmetric positive definite matrices,

$$\widehat{s}_M \in \operatorname*{argmin}_{g \in \mathcal{H}^p} \left\{ \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(Y_i^j - g^j(X_i)\right)^2 + \sum_{j=1}^{p} \sum_{\ell=1}^{p} M_{j,l} \left\langle g^j, g^\ell \right\rangle_{\mathcal{H}} \right\}. \tag{6.8}$$

For instance, taking

$$M = \frac{1}{p} \operatorname{diag}(\lambda_1, \dots, \lambda_p)$$

in Eq. (6.8) corresponds to using kernel ridge regression separately over the p tasks, and taking

$$M = (\lambda + p\mu) I_p - \mu \mathbf{1}\mathbf{1}^\top$$

in Eq. (6.8) leads to the estimator of Eq. (6.7). The remaining problem is to choose M from data, which is an estimator selection problem.

Let us now consider the fixed-design (multivariate) regression framework, as in Section 4.2. Then, defining $F_i^j = \eta^j(X_i)$ and writing the $n \times p$ matrices

$$\big(Y_i^j\big)_{1\le i,j\le n}, \qquad \big(F_i^j\big)_{1\le i,j\le n} \qquad \text{and} \qquad \big(\varepsilon_i^j\big)_{1\le i,j\le n}$$

as vectors Y, F, $\varepsilon \in \mathbb{R}^{np}$ by stacking their columns, we can summarize the problem as follows. We observe

$$Y = F + \varepsilon$$

and the goal is to find, from the data Y only, some $t \in \mathbb{R}^{np}$ such that its quadratic excess loss

$$\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\big(t_i^j - F_i^j\big)^2 = \frac{1}{np}\|t - F\|^2$$

is small.

For any $M \in \mathcal{S}_p^{++}$, let $\widehat{F}_M = (\widehat{s}_M^j(X_i))_{i,j}$ denote the kernel ridge estimator induced by Eq. (6.8) in the fixed-design framework. By the representer's theorem, **[9]** proves that

$$\widehat{F}_M = A_M Y \quad \text{where} \quad A_M = \big(M^{-1}\otimes K\big)\Big(\big(M^{-1}\otimes K\big) + npI_{np}\Big)^{-1}$$

hence $\widehat{F}_M$ is a linear estimator, as the estimators we consider in Section 4.6 for univariate regression.

Assume that a family $(\widehat{F}_M)_{M\in\mathcal{M}}$ is given, for some $\mathcal{M} \subset \mathcal{S}_p^{++}$. The penalization approach consists in choosing

$$\widehat{M} \in \operatorname*{argmin}_{M\in\mathcal{M}}\left\{\frac{1}{np}\left\|\widehat{F}_M - F\right\|^2 + \operatorname{pen}(M)\right\}$$

for some penalty function $\operatorname{pen} : \mathcal{M} \to \mathbb{R}$. The analysis of Section 4.6 can be extended to multivariate regression, leading to the optimal penalty

$$\operatorname{pen}_{\mathrm{opt}}(M) = \frac{2\operatorname{tr}\big(A_M\cdot(\Sigma\otimes I_{np})\big)}{np}\ . \tag{6.9}$$

Compared to the univariate case—Eq. (4.30) in Section 4.6—, the optimal penalty here depends on the unknown covariance matrix $\Sigma \in \mathcal{S}_p^+$, instead of being known up to some multiplicative factor σ^2 only.

The approach of **[9]** for this estimator selection problem is in two steps: (i) find some estimator $\widehat{\Sigma}$ of the covariance matrix Σ, and (ii) plug $\widehat{\Sigma}$ into the penalty (6.9) instead of Σ. The first step relies heavily on the minimal penalty algorithm of **[18]**. As explained in Section 4.6, for any (univariate) regression problem, one can estimate the residual variance under mild assumptions. Then, if $(e_1,\dots,e_n)$ denotes the canonical basis of $\mathbb{R}^n$, projecting the multi-task problem (6.6) onto e_i allows to estimate $\Sigma_{i,i}$, and projecting the multi-task problem (6.6) onto $e_i + e_j$ with $i \neq j$ allows to estimate

$\Sigma_{i,i} + \Sigma_{j,j} + 2\Sigma_{i,j}$. Gathering together these variance estimators yields an estimator $\widehat{\Sigma}$, which is proved in [9] to be close to Σ on a large probability event, with non-asymptotic multiplicative deviation terms proportional to

$$p\sqrt{\frac{\log(n)}{n}}c(\Sigma)^2$$

where $c(\Sigma)$ is the condition number of Σ, that is, the ratio between its largest and smallest eigenvalues. In particular, $\widehat{\Sigma}$ consistently estimates Σ even in asymptotic settings where the number of tasks p grows with n, provided it does not grow too fast. Note that a tighter bound is obtained if Σ is known to be diagonal, allowing p to grow as fast as n^δ for any $\delta > 0$.

A first-order optimal oracle inequality (2.2) is proved in [9] for the resulting estimation procedure, with deviation terms small enough if p and $c(\Sigma)$ are not too large compared to n. Experiments on synthetic data show that this can lead to better performance compared to considering separately the p regression tasks; see also [Sol13] for a more detailed analysis of when the multi-task approach can lead to such an improvement.

6.4. Margin adaptivity in classification

One of the major challenges of estimator selection is adaptivity to unknown properties of the data. In the binary classification framework with the 0–1 loss—see Example 2.3—, usual minimax learning rates are of order $1/\sqrt{n}$ at least [BBL05, Section 5.5]. Faster rates—up to $1/n$—can nevertheless be obtained under the margin condition [MT99], that is, if for some $\varepsilon_0, C_0 > 0$ and $\alpha \geq 1$,

$$\forall \varepsilon \in (0, \varepsilon_0], \qquad \mathbb{P}\Big(\big|2\eta(X) - 1\big| \leq \varepsilon\Big) \leq C_0 \varepsilon^\alpha \ .$$

The extreme situation "$\alpha = +\infty$" corresponds to assuming

$$\mathbb{P}\Big(\big|2\eta(X) - 1\big| \leq h\Big) = 0 \tag{6.10}$$

for some $h > 0$, which makes the classification problem easier since it provides an upper bound on the noise level. For simplicity, from now on we focus on (6.10), see [8] for the general case.

Following the approach of Koltchinskii [Kol06], the margin condition (6.10) can be replaced by

$$\forall t \in \mathbb{S}, \qquad \ell(s^\star, t) \geq h \operatorname{var}_P\big(\gamma(t;\cdot) - \gamma(s^\star;\cdot)\big) \tag{6.11}$$

for some $h > 0$, which is implied by Eq. (6.10) and leads to the same fast rates. Let us now consider the model selection problem among a given family $(S_m)_{m\in\mathcal{M}}$. For each $m \in \mathcal{M}$, risk bounds for the corresponding empirical risk minimizer $\widehat{s}_m$ can be obtained with the same fast rates under the weaker condition

$$\forall t \in S_m, \qquad \ell(s^\star, t) \geq h_m \operatorname{var}_P\big(\gamma(t;\cdot) - \gamma(s^\star;\cdot)\big) \ , \tag{6.12}$$

see [MN06]. Since Eq. (6.12) only involves classifiers t in S_m, we call it a "local" margin condition.

For instance, assuming every model S_m has finite VC-dimension $V_m \geq 1$ and Eq. (6.12) holds true, the empirical risk minimizer $\widehat{s}_m$ on S_m satisfies

$$\mathbb{E}[\ell(s^\star, \widehat{s}_m)] \leq C\left(\ell(s^\star, S_m) + \min\left\{\frac{\log(n)V_m}{nh_m}\,,\, \sqrt{\frac{V_m}{n}}\right\}\right)$$

for some numerical constant $C > 0$. Minimax lower bounds show that this cannot be improved in general [MN06]. Then, for any model selection procedure $\widehat{m}$, we cannot hope to prove an oracle inequality better than

$$\ell(s^\star, \widehat{s}_{\widehat{m}}) \leq C \inf_{m\in\mathcal{M}}\left\{\ell(s^\star, S_m) + \min\left\{\frac{\log(n)V_m}{nh_m}\,,\, \sqrt{\frac{V_m}{n}}\right\}\right\}\,. \tag{6.13}$$

Moreover, proving that Eq. (6.13) holds for some procedure $\widehat{m}$ not using the knowledge of $(h_m)_{m\in\mathcal{M}}$ is a real challenge, that we call "strong margin adaptivity". In particular, this is more challenging than adapting to the "global" margin condition (6.11), since Eq. (6.12) can hold for some models S_m with h_m much larger than the "uniform" margin h for which Eq. (6.11) holds true [8].

Two main results are proved in [8] about this problem. First, when the models S_m are nested, strong margin adaptivity (6.13) holds for penalization with local Rademacher complexities [BMP04, LW04, BBM05, Kol06]. Note that local Rademacher complexities are resampling-based penalties—as the ones considered in Section 3.1—which do not estimate $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)]$ but some upper bound on it. Compared to previous results, in particular the ones of [Kol06], the oracle inequality in [8] is the first of the form (6.13) to be proved for a procedure not using the knowledge of $(h_m)_{m\in\mathcal{M}}$.

Second, [8] proves strong margin adaptivity is not always possible. A family of (non-nested) models is built such that, for every sample size n and every model selection procedure $\widehat{m}$, a distribution P exists for which, with a positive probability, $\widehat{m}$ fails to be strongly margin adaptive. Nevertheless, this is only a worst-case lower bound, and the nested assumption is not fully necessary: [8] also shows some specific situations where strong margin adaptivity is possible without having nested models.

6.5. Multi-label classification

Multi-label classification corresponds to Example 2.3 in Section 2.1.1 when the variable of interest is a subset of a given set $\mathcal{V}$ of possible labels. For instance, in image, video [XHE+10] or text [Joa98] tagging, a given image, text or video can be assigned several labels simultaneously. Then, $\mathcal{Y} = \mathfrak{P}(\mathcal{V})$ has cardinality $2^{\mathrm{Card}(\mathcal{V})}$, which is huge unless $\mathcal{V}$ is very small. So, an exhaustive search over $\mathcal{Y}$ is not possible and standard multiclass classification algorithms cannot be used.

A standard approach to multi-label classification is "one versus rest" (OvR): for every $v \in \mathcal{V}$, build a predictor $\widehat{s}_v$ for the associated binary classification problem—for

every $x \in \mathcal{X}$, $\widehat{s}_v(x)$ answers the question "Is v present among the labels of x?"—and define

$$\widehat{s}(x) = \left\{ v \in \mathcal{V} \,/\, \widehat{s}_v(x) = 1 \right\}.$$

The drawback of OvR is that it does not take into account possible relationships between labels. For instance, for images, 'zebra' and 'lion' are more likely to occur together than 'zebra' and 'reindeer', an information which can be learned from data in order to improve the classification performance on new examples.

Remark that multi-label classification can be seen as a multi-task problem—see Section 6.3—where the tasks are the binary classification problems corresponding to each label $v \in \mathcal{V}$ separately. Then, OvR is the usual single-task approach. Compared to the framework of Section 6.3, a major difference lies in what makes the tasks "similar". Here, tasks are similar because of their dependence structure, that is, Σ with the notation of Section 6.3: some labels often or seldom occur together. In Section 6.3, the assumed similarity is between the regression functions $\eta^1, \ldots, \eta^p$.

A new procedure for multi-label classification is proposed in **[21]**, which can learn attractive and repulsive priors among labels while being computationally tractable. When the prior among labels is given, multi-label classification as in **[21]** leads to a rather standard classifier. The challenge is to be able to learn the prior from data in a computationally efficient way, which can be seen as an estimator selection problem. The main lines of **[21]** share some similarities with **[12]** and **[13]**. Since multi-label classification is a structured prediction problem, the large-margin relaxation of [**THJA05**] can be used, and what remains—which is not an easy task, even if details are omitted here—is to make this relaxation tractable for potentially large sets $\mathcal{V}$. Experiments on real data sets—with $\mathrm{Card}(\mathcal{V})$ up to 159—show that incorporating these priors can improve over existing methods, which take no prior into account, or only attractive priors.

CHAPTER 7

Prospects

This chapter describes some directions for future work on estimator selection procedures in statistical learning, related to the results shown in Chapters 2–6.

My main concern remains the one presented at the beginning of Chapter 2: provide theoretical results that can help practitioners, and more specifically:

(1) explain the differences that can be observed empirically between several procedures,

(2) fully take into account the computational complexity when analyzing some procedures, in order to solve computational trade-offs,

(3) provide theoretical grounds for widely-used methods that are known to work well in practice, and

(4) propose new algorithms that can be useful in practice, for instance by taking into account the underlying structure of high-dimensional data (as with multi-task learning in Section 6.3) or by allowing to analyze data of complex nature such as videos, DNA sequences or graphs (following Section 5.3 for instance).

Following Chapters 2–6, we present first some research directions for taking into account second-order terms in the general estimator selection problem (Section 7.1), then for cross-validation and resampling methods (Section 7.2), for minimal penalty algorithms (Section 7.3), and finally for change-point detection procedures (Section 7.4).

7.1. Second-order terms in the comparison of estimator selection methods

Section 2.2.5 provides a heuristics as a first step for taking into account the "variance" of the criterion $\mathcal{C}(m)$ when analyzing an estimator selection procedure of the form of Eq. (2.4), that is,

$$\widehat{m}_{\mathcal{C}} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathcal{C}(m) \right\} \ .$$

This heuristics already provides good results on synthetic data for V-fold cross-validation (VFCV) procedures [14]. Several questions then arise: (i) Can this heuristics be formalized, at least in some specific frameworks? (ii) Can this heuristics lead to a *quantitative* comparison of estimator selection procedures which differ only through their variances? Some partial answer to the second question is already provided at the end of Section 2.2.5, but it needs to be made more precise, in particular for being able to validate it on synthetic data.

Procedure	L–Dya2	S–Dya2
pen2F	10.21 ± 0.08	2.39 ± 0.01
pen5F	7.47 ± 0.06	2.16 ± 0.01
pen10F	6.89 ± 0.06	2.11 ± 0.01
penLOO	6.35 ± 0.05	2.06 ± 0.01
2FCV	6.41 ± 0.05	2.05 ± 0.01
5FCV	6.27 ± 0.05	2.05 ± 0.01
10FCV	6.24 ± 0.05	2.05 ± 0.01
LOO	6.34 ± 0.05	2.06 ± 0.01

TABLE 7.1. Extract from Table 2 in [14]: estimated model selection performance $\mathbb{E}[\ell(s^\star, \widehat{s}_{\widehat{m}})/\inf_{m\in\mathcal{M}} \ell(s^\star, \widehat{s}_m)]$ for several V-fold penalization (top) and V-fold cross-validation (bottom) model selection procedures $\widehat{m}$ in least-squares density estimation.

An important example is the comparative analysis of VFCV procedures. A theoretical analysis of their variance as a function of V is provided in [14] for projection estimators in least-squares density estimation. Some work in progress, in collaboration with Matthieu Lerasle and Nelo Magalhães, suggests that this analysis can be generalized to kernel density estimators, which would have a large impact since it includes the bandwidth choice problem. Looking for other frameworks where the variance of VFCV criteria can be theoretically assessed is also an important research direction for at least two reasons. First, as reported in [7], empirical results suggest the dependence on V of the variance of VFCV is not the same in all frameworks. The conclusions of [14] certainly are still valid for least squares and similar methods, but identifying precisely some estimators for which different behaviors can arise would have a wide impact. Second, specificities of the least-squares density estimation made possible to analyze VFCV in [14] at a level of precision never attained before in any other framework. Other settings might allow even more precise theoretical analyses, in particular by turning the heuristics of Section 2.2.5 into rigorous quantitative statements.

Other second-order terms should be taken into account when analyzing procedures of the form of Eq. (2.4). For instance, let us consider Table 7.1 where some experimental results from [14] are reported. The performance of V-fold penalization procedures improves when V increases from $V = 2$ to $V = n$ (penLOO), as predicted by the theoretical results of [14]. On the contrary, the performance of VFCV procedures as a function of V seems difficult to analyze. According to the results of Chapter 3, when V increases, the performance should improve at first order (because the bias of VFCV decreases) and at second order (because the variance of VFCV decreases). In Table 7.1, a completely different behavior is observed: in setting 'L–Dya2', $V = 10$ gives the best

performance, significantly better than $V = n = 500$, and in setting 'S–Dya2', all V yield approximately the same performance.

Analyzing more precisely the experiments of **[14]** shows that in the settings considered there, the performance of estimator selection procedures is better when the criterion $\mathcal{C}(m)$ is slightly biased as an estimator of $\mathbb{E}[\mathcal{R}(m)]$. For penalization procedures defined by Eq. (4.1), this means that a smaller final risk is obtained thanks to *overpenalization* by a well-chosen factor. For instance, the risk obtained with the (theoretical) penalty $C \times \mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)]$ is not optimal for $C = 1$—as suggested by first-order theoretical results and by the unbiased risk estimation principle of Section 2.2.2—but for $C \approx 2$ in setting 'L–Dya2' and $C \approx 1.5$ in setting 'S–Dya2'. A similar phenomenon is well known in practice—it is better to make a slightly "too conservative" model choice—and it has been observed in other frameworks, as for instance in **[16]** in regression.

Although the heuristics of Section 2.2.5 is formulated for taking into account the variance, it can actually also be used in order to explain the overpenalization phenomenon. For instance, for comparing $\mathcal{C}_1$ and $\mathcal{C}_2$ that do not have the same expectation, Eq. (2.16)–(2.17) are still valid, so it remains true that if $\mathrm{SNR}_{\mathcal{C}_1}(m) > \mathrm{SNR}_{\mathcal{C}_2}(m)$ for all $m \neq m^\star$, $\mathcal{C}_1$ should be better than $\mathcal{C}_2$. Let us assume $\mathcal{C}_1$ and $\mathcal{C}_2$ are two deterministic penalization criteria of the form

$$\mathcal{C}_i(m) = P_n\gamma(\widehat{s}_m) + \kappa_i\,\mathbb{E}\big[\mathrm{pen}_{\mathrm{id}}(m)\big]$$

for some constants $\kappa_1 \neq \kappa_2$. Then, the variance of $\Delta_{\mathcal{C}_i}(m, m') = \mathcal{C}_i(m) - \mathcal{C}_i(m')$ does not depend on i. We do not know currently how to go further because when $\kappa_1 \neq \kappa_2$, some $m, m' \in \mathcal{M}$ exist such that $\mathrm{SNR}_{\mathcal{C}_1}(m) > \mathrm{SNR}_{\mathcal{C}_2}(m)$ and $\mathrm{SNR}_{\mathcal{C}_1}(m') < \mathrm{SNR}_{\mathcal{C}_2}(m')$. Making the heuristics of Section 2.2.5 quantitative would help point out which elements m of $\mathcal{M}$ matter most for the final risk of $\widehat{m}_{\mathcal{C}_i}$, hence making possible to "guess" the optimal overpenalizing constant $C^\star$. Even without a full theoretical justification, such a guess that works well empirically could have a high practical impact.

7.2. Cross-validation and resampling methods

Several research directions on cross-validation (CV) methods have been detailed in Section 7.1. Let us mention here two other open problems on these methods.

First, V-fold penalization is proved to correct for the bias of VFCV in two settings—regressograms and projection estimators in least-squares density estimation—, leading to first-order optimal oracle inequalities, but VFCV is widely used, far beyond these two settings. Therefore, extending the theoretical analysis of V-fold penalties would be of practical interest, since they allow to correct for the bias or to overpenalize easily—see Section 7.1—, depending on the problem at hand. For instance, empirical results suggest that this is possible for support vector regression and for CART estimators in regression [**DBCU14**].

Second, when the computational cost is a key issue—for instance in the "big data" framework—the specific learning and optimization algorithms that have been developed

rely on some hyperparameters. CV methods are natural candidates for choosing them, but then we must carefully take into account the computational cost they induce. For instance, we cannot hope to be able to compute the 10-fold cross-validation estimator of the risk of these algorithms for a large set of values of their parameters, and then choose values of hyperparameters which minimize the estimated risk. Recent works proposed to modify resampling methods for computational reasons, for instance [**KTSJ12, Kuh14**]. Nevertheless, these works do not assess precisely the statistical performance of the resulting procedure in terms of estimator selection, and a precise non-asymptotic analysis such as the ones of Chapter 3 remains to be done in this framework.

7.3. Minimal penalty algorithms

Gathering recent theoretical results on minimal penalties in the survey paper **[22]** leads to sketching a set of frameworks where similar ideas can be extended. More precisely, the key quantities for the theoretical study of minimal penalties—as in Sections 4.4–4.5—are, for every $m \in \mathcal{M}$,

$$p_1(m) := P\bigl(\gamma(\widehat{s}_m) - \gamma(s_m^\star)\bigr) \quad \text{and} \quad p_2(m) := P_n\bigl(\gamma(s_m^\star) - \gamma(\widehat{s}_m)\bigr)$$

where $s_m^\star \in \operatorname{argmin}_{t \in S_m} \ell(s^\star, t)$ in the model selection setting, and $s_m^\star$ is a well-chosen fixed element of $\mathbb{S}$ in the general case. If precise concentration inequalities can be obtained for p_1 and p_2, and if their expectations can be compared, theoretical guarantees can be obtained on the resulting minimal penalty algorithm. Alternatively, if

$$\left| \frac{p_1(m) - p_2(m)}{p_1(m)} \right|$$

is small on a large probability event, the slope heuristics $\mathrm{pen}_{\mathrm{opt}} \approx 2\,\mathrm{pen}_{\mathrm{min}}$ can be validated, up to (usually minor) technical issues, see **[22]**. For instance, it seems that minimal penalty algorithms can be developed for choosing among kernel density estimators [**LMaRB14**], or among linear estimators in heteroscedastic regression. Note that having $p_1 \approx p_2$ can be linked with the so-called Wilks phenomenon [**BM11**].

An important open problem remains the case of "large" families of estimators, as defined in Section 2.2.3. Only one weak theoretical result on minimal penalties is available in such a framework [**BM07**], but numerical experiments suggest that the slope heuristics still works well more generally, for instance for change-point detection [**Leb05**]. More precisely, it seems that minimal penalty algorithms automatically *adapt to the richness* of $(\widehat{s}_m)_{m \in \mathcal{M}}$, which is possible since Algorithm 2 makes use of the full family $(\widehat{s}_m)_{m \in \mathcal{M}}$ for building a constant $\widehat{C}$ such that $2\widehat{C}\,\mathrm{pen}_1$ is close to an optimal penalty.

The latter property suggests another practical application of minimal penalty algorithms. Indeed, there is a continuum between "small" and "large" families of estimators, according to the terminology of Section 2.2. If minimal penalties adapt to large families and are proved to be first-order optimal for small families, we can expect them to

overpenalize a bit in the middle. From a non-asymptotic perspective, the boundary between small and large families is unclear, and when the signal-to-noise ratio is not large enough, an asymptotically "small" family may look like if it was "large". Hence, minimal penalty algorithms might provide an automatic way to overpenalize when necessary, an important practical problem that we discuss in Section 7.1. Experimental results in [9] support this conjecture, by showing examples where a smaller risk is obtained by penalizing with the data-driven penalty

$$\frac{2\operatorname{tr}\Big(A_M \cdot \big(\widehat{\Sigma} \otimes I_{np}\big)\Big)}{np}$$

than with the (theoretical) optimal penalty

$$\frac{2\operatorname{tr}\big(A_M \cdot (\Sigma \otimes I_{np})\big)}{np} \ .$$

More details about the setting of [9] are provided in Section 6.3.

7.4. Change-point detection

Change-point detection in high-dimensional time series, or when observations are structured objects such as histograms, strings or graphs, is an important problem for many application domains, as explained in Chapter 5. In particular, Section 5.3 describes a procedure proposed in [19] that can tackle such a problem thanks to the combination of a model selection approach with kernel methods. Several practical problems for using this procedure are still widely open, and we discuss below three main issues that should be addressed.

First, in the Hilbertian setting that must be considered for analyzing the kernel change-point procedure of [19], data are not homoscedastic. Although the theoretical results in [19] take this fact into account, the first step of the procedure remains empirical risk minimization for every fixed number of change-points, and [6] shows that in the one-dimensional case this can lead to overfitting. A natural idea would be to combine the kernel change-point approach of [19] with the cross-validation based algorithm of [6]. Although writing down the resulting procedure is rather straightforward, making it computationally tractable and validating it—at least empirically—remains to be done.

Second, the properties of the kernel change-point detection procedure of [19] heavily depends on the choice of a kernel k. For instance, for one-dimensional data, the linear kernel $k(x, y) = xy$ leads to the procedure of [Leb05] which detects changes in the mean of the signal—assuming the variance is constant—but not in other moments of the distribution. On the contrary, a Gaussian kernel

$$k(x, y) = \exp\left(\frac{-(x-y)^2}{2h^2}\right)$$

can make the procedure detect changes in the distribution of the signal even when the mean and variance are constant, see [19].

More generally, given a kernel k, can we describe theoretically which kind of changes in the distribution of the Y_i the procedure of [19] can detect? A first step of the analysis might be the oracle inequality proved in [19]: it shows that the procedure is (close to) optimal in terms of risk in the reproducing kernel Hilbert space $\mathcal{H}$ associated with k. So, depending on k, a given change in the distribution of Y_i may or may not induce a large value of the risk in $\mathcal{H}$ for all estimators which do not detect the corresponding change-point. Therefore, if k is well-chosen, an oracle inequality such as the one of [19] should imply a consistent identification of the change-points locations.

Conversely, if we can describe theoretically which kind of change in the distribution we are looking for—for instance, a change in the third moment for one-dimensional data, or a change in the correlation matrix for multivariate data—, which kernel should we use in the procedure of [19]? A first (rough) answer might come from the understanding of what the procedure is doing for the most classical kernels. Let us also mention an interesting advantage of the cross-validation/kernel change-point detection procedure suggested above, result of the combination of [6] and [19]. In the one-dimensional case, [6] allows to detect changes in the mean of the distribution, whatever the variance and other moments of the distribution, which can change anytime without perturbing the procedure. In combination with kernels, we can expect the resulting procedure to detect changes in the mean-element of the distribution, whatever its other parameters. Then, if we could design a kernel such that the mean-element contains exactly the features of the distribution we are interested in, we would be able to detect exactly changes in these features, allowing other aspects of the distribution to change.

Third, it is often not realistic to assume that we can describe theoretically which kind of change in the distribution we are interested in. For instance, how to describe non-trivial distribution changes in a series of strings, or in a series of graphs? A more realistic option is to ask an expert of the problem considered to provide some examples of segmented time series, assuming they are similar to the time series we want to automatically segment. Then, if $\mathcal{Y} = \mathbb{R}^d$, learning the best kernel among linear kernels $k_A(x, y) = \langle Ax, Ay \rangle$, $A \in \mathcal{M}_d(\mathbb{R})$, exactly corresponds to the metric learning problem considered in Section 5.4. If we could extend this approach to more general families of kernels, we would obtain a computationally efficient and principled data-driven procedure for learning the best kernel to be used in the procedure of [19].

Bibliographie exogène

[Aka73] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[All74] David M. Allen. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125–127, 1974.

[Aud04] Jean-Yves Audibert. A better variance control for pac-bayesian classification. Technical Report 905b, Laboratoire de Probabilités et Modèles Aléatoires, 2004.

[BA02] Kenneth P. Burnham and David R. Anderson. *Model Selection and Multimodel Inference*. Springer-Verlag, New York, second edition, 2002. A practical information-theoretic approach.

[BB08] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20, pages 161–168. NIPS Foundation (http://books.nips.cc), 2008.

[BBL05] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375 (electronic), 2005.

[BBM99] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[BBM05] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 2005.

[BD93] Boris E. Brodsky and Boris S. Darkhovsky. *Nonparametric methods in change-point problems*, volume 243 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1993.

[BDL08] Gérard Biau, Luc P. Devroye, and Gábor Lugosi. Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.*, 9:2015–2033, 2008.

[BFOS84] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA, 1984.

[BGH10] Yannick Baraud, Christophe Giraud, and Sylvie Huet. Estimator selection in the gaussian setting. *Ann. Inst. H. Poincaré Probab. Statist.*, 2010. To appear. arXiv:1007.2096.

[Bia12] Gérard Biau. Analysis of a random forests model. *J. Mach. Learn. Res.*, 13:1063–1095, 2012.

[BKL+09] Leif Boysen, Angela Kempe, Volkmar Liebscher, Axel Munk, and Olaf Wittich. Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Statist.*, 37(1):157–183, 2009.

[BM97] Lucien Birgé and Pascal Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[BM01] Lucien Birgé and Pascal Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001.

[BM06] Gilles Blanchard and Pascal Massart. Discussion: "Local Rademacher complexities and oracle inequalities in risk minimization" [Ann. Statist. **34** (2006), no. 6, 2593–2656] by V. Koltchinskii. *Ann. Statist.*, 34(6):2664–2671, 2006.

[BM07] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

[BM11] Stéphane Boucheron and Pascal Massart. A high dimensional wilks phenomenon. *Probab. Theory Related Fields*, 150(3-4):405–433, 2011.

[BMM11] Jean-Patrick Baudry, Cathy Maugis, and Bertrand Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, pages 1–16, 2011. 10.1007/s11222-011-9236-1.

[BMP04] Peter L. Bartlett, Shahar Mendelson, and Petra Philips. Local complexities for empirical risk minimization. In *Learning theory*, volume 3120 of *Lecture Notes in Comput. Sci.*, pages 270–284. Springer, Berlin, 2004.

[BN93] Michèle Basseville and Igor V. Nikiforov. *Detection of abrupt changes: theory and application*. Prentice Hall Information and System Sciences Series. Prentice Hall Inc., Englewood Cliffs, NJ, 1993.

[Bre00] Leo Breiman. Some infinity theory for predictor ensembles. Technical Report Technical Report 577, U.C. Berkeley Department of Statistics, August 2000. available at http://www.stat.berkeley.edu/tech-reports/577.pdf.

[Bre01] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[BS92] Leo Breiman and Philip Spector. Submodel Selection and Evaluation in Regression. The X-Random Case. *International Statistical Review*, 60(3):291–319, 1992.

[Bur89] Prabir Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514, 1989.

[Bur90] Prabir Burman. Estimation of optimal transformations using v-fold cross validation and repeated learning-testing methods. *Sankhyā Ser. A*, 52(3):314–345, 1990.

[BV11] Kevin Bleakley and Jean-Philippe Vert. The group fused lasso for multiple change-point detection. Technical report, arXiv, 2011. arXiv:1106.4199.

[BvdG11] Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

[Cat07] Olivier Catoni. *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes Monograph Series*. Inst. Math. Statist., 2007.

[Cel08] Alain Celisse. *Model Selection Via Cross-Validation in Density Estimation, Regression and Change-Points Detection*. PhD thesis, University Paris-Sud 11, December 2008. oai:tel.archives-ouvertes.fr:tel-00346320_v1.

[CH06] Thomas M. Cover and Peter E. Hart. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theor.*, 13(1):21–27, September 2006.

[CJ13] Venkat Chandrasekaran and Michael I. Jordan. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 110(13):E1181–E1190, 2013.

[CR04] Fabienne Comte and Yves Rozenholc. A new algorithm for fixed design regression and denoising. *Ann. Inst. Statist. Math.*, 56(3):449–473, 2004.

[CSS+07] Arshia Cont, Diemo Schwarz, Norbert Schnell, Christopher Raphael, et al. Evaluation of real-time audio-to-score alignment. In *Proceedings ISMIR*, 2007.

[DAV10] Josselin Desmars, Jean-Eudes Arlot, and Alain Vienne. Estimation of accuracy of close encounter performed by the bootstrap method. *Cosmic Research*, 48:472–478, October 2010.

[DBCU14] Charanpal Dhanjal, Nicolas Baskiotis, Stéphan Clémençon, and Nicolas Usunier. An empirical comparison of v-fold penalisation and cross-validation for model selection in distribution-free regression. *Pattern Analysis and Applications*, pages 1–13, 2014.

[Des09] Josselin Desmars. *Précision d'extrapolation des éphémérides des objets du système solaire. Application aux satellites de Saturne.* These, Observatoire de Paris, June 2009.

[DGL96] Luc P. Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

[DW79] Luc P. Devroye and Terry J. Wagner. Distribution-Free performance Bounds for Potential Function Rules. *IEEE Transaction in Information Theory*, 25(5):601–604, 1979.

[Efr79] Bradley Efron. Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 1979.

[Efr83] Bradley Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78(382):316–331, 1983.

[EMP05] Theodoros Evgeniou, Charles A. Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *J. Mach. Learn. Res.*, 6:615–637, 2005.

[FH51] Evelyn Fix and Joseph L. Hodges. Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical Report 21-49-004,4, U.S. Air Force, School of Aviation Medicine, Randolph Field, TX, 1951.

[FH89] Evelyn Fix and Joseph L. Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review*, 57(3):238–247, 1989.

[Gei75] Seymour Geisser. The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70:320–328, 1975.

[Gen10] Robin Genuer. *Forêts aléatoires: aspects théoriques, sélection de variables et applications.* PhD thesis, University Paris-Sud 11, 2010. http://tel.archives-ouvertes.fr/tel-00550989/.

[Gen12] Robin Genuer. Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, 24(3):543–562, 2012.

[Gir14] Christophe Giraud. *Introduction to High-Dimensional Statistics.* Monographs on Statistics and Applied Probability. Chapman and Hall/CRC, Boca Raton, FL, 2014. Forthcoming.

[GKKW02] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-free Theory of Nonparametric Regression.* Springer Series in Statistics. Springer-Verlag, New York, 2002.

[HO93] Per Christian Hansen and Dianne Prost O'Leary. The use of the *L*-curve in the regularization of discrete ill-posed problems. *SIAM J. Sci. Comput.*, 14(6):1487–1503, 1993.

[HRB03] Christian Houdré and Patricia Reynaud-Bouret. Exponential inequalities, with constants, for U-statistics of order two. In *Stochastic inequalities and applications*, volume 56 of *Progr. Probab.*, pages 55–69. Birkhäuser, Basel, 2003.

[HTF01] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer-Verlag, New York, 2001. Data mining, inference, and prediction.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.

[HVLYFC09] Zaïd Harchaoui, Félicien Vallet, Alexandre Lung-Yut-Fong, and Olivier Cappé. A Regularized Kernel-based Approach to Unsupervised Audio Segmentation. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.

[Joa98] Thorsten Joachims. Text categorization with Support Vector Machines: Learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Machine Learning: ECML-98*, volume 1398 of *Lecture Notes in Computer Science*, pages 137–142. Springer Berlin Heidelberg, 1998.

[Kol06] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 2006.

[KTSJ12] Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I. Jordan. The big data bootstrap. In *29th International Conference on Machine Learning (ICML 2012)*, 2012. arXiv:1206.6415.

[Kuh14] Max Kuhn. Futility analysis in the cross-validation of machine learning models, May 2014. arXiv:1405.6974.

[Lav05] Marc Lavielle. Using penalized contrasts for the change-point problem. *Signal Proces.*, 85(8):1501–1510, 2005.

[Leb05] Émilie Lebarbier. Detecting multiple change-points in the mean of a Gaussian process by model selection. *Signal Proces.*, 85:717–736, 2005.

[Ler11] Matthieu Lerasle. Optimal model selection for stationary data under various mixing conditions. *Ann. Statist.*, 39(4):1852–1877, 2011.

[Ler12] Matthieu Lerasle. Optimal model selection in density estimation. *Ann. Inst. Henri Poincaré Probab. Stat.*, 48(3):884–908, 2012.

[LMaRB14] Matthieu Lerasle, Nelo Magalhães, and Patricia Reynaud-Bouret. Optimal kernel selection for density estimation. Private communication, 2014.

[LT06] Marc Lavielle and Gilles Teyssière. Detection of Multiple Change-Points in Multivariate Time Series. *Lithuanian Mathematical Journal*, 46:287–306, 2006.

[LW04] Gábor Lugosi and Marten Wegkamp. Complexity regularization via localized random penalties. *Ann. Statist.*, 32(4):1679–1697, 2004.

[Mal73] Colin L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

[Mas07] Pascal Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[MM11] Cathy Maugis and Bertrand Michel. A non asymptotic penalized criterion for gaussian mixture model selection. *ESAIM Probab. Stat.*, 15:41–68, 2011.

[MN06] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, 2006.

[MT99] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 1999.

[Nad64] Elizbar A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9(1):141–142, 1964.

[Pic05] Franck Picard. *Process segmentation/clustering Application to the analysis of array CGH data*. PhD thesis, Université Paris-Sud 11, 2005. http://tel.archives-ouvertes.fr/tel-00116025/fr/.

[PLBR11] Franck Picard, Émilie Lebarbier, E. Budinska, and Stéphane Robin. Joint segmentation of multivariate gaussian processes using mixed linear models. *Comput. Statist. Data Anal.*, 55(2):1160–70, 2011.

[Rig10] Guillem Rigaill. Pruned dynamic programming for optimal multiple change-point detection. Technical report, arXiv, April 2010. arXiv:1004.0887.

[Sau10a] Adrien Saumard. *Estimation par Minimum de Contraste Régulier et Heuristique de Pente en Sélection de Modèles*. PhD thesis, Université de Rennes 1, October 2010. http://tel.archives-ouvertes.fr/tel-00569372/fr/.

[Sau10b] Adrien Saumard. Nonasymptotic quasi-optimality of AIC and the slope heuristics in maximum likelihood estimation of density using histogram models. hal-00512310, September 2010.

[SBV14] Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests, May 2014. arXiv:1405.2881.

[SC08] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008.

[Sch78] Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

[Sha93] Jun Shao. Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, 88(422):486–494, 1993.

[Shi97] Ritei Shibata. Bootstrap estimate of Kullback-Leibler information for model selection. *Statist. Sinica*, 7(2):375–394, 1997.

[Sol13] Matthieu Solnon. Comparison between multi-task and single-task oracle risks in kernel ridge regression, July 2013. arXiv:1307.5286.

[SS01] Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[Sto74] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147, 1974. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors.

[THJA05] Ioannis Tsochantaridis, Thomas Hoffman, Thorsten Joachims, and Yosemin Altun. Support vector machine learning for interdependent and structured output spaces. *J. Mach. Learn. Res.*, 6(Sep):1453–1484, 2005.

[TNM14] Alexander Tartakovsky, Igor Nikiforov, and Basseville Michèle. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*, volume 136 of *Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC, Boca Raton, FL, 2014.

[TPP99] Julie D. Thompson, Frédéric Plewniak, and Olivier Poch. Balibase: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, 15(1):87–88, 1999.

[Vap82] Vladimir Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer Series in Statistics. Springer-Verlag, New York, 1982. Translated from the Russian by Samuel Kotz.

[vD03] Mark J. van der Laan and Sandrine Dudoit. Unified Cross-Validation Methodology For Selection Among Estimators and a General Cross-Validated Adaptive Epsilon-Net Estimator: Finite Sample Oracle Inequalities and Examples. Working Paper Series Working Paper 130, U.C. Berkeley Division of Biostatistics, November 2003. available at http://www.bepress.com/ucbbiostat/paper130.

[Wah90] Grace Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

[Was06] Larry Wasserman. *All of nonparametric statistics*. Springer Texts in Statistics. Springer, New York, 2006.

[Wat64] Geoffrey Stuart Watson. Smooth regression analysis. *Sankhyā Ser. A*, 26:359–372, 1964.

[XHE+10] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3492, June 2010.

[YA89] Yi-Ching Yao and Siu-Tong Au. Least-squares estimation of a step function. *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)*, 51(3):370–381, 1989.

[Yan05] Yuhong Yang. Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005.

[Yan07] Yuhong Yang. Consistency of cross validation for comparing regression procedures. *Ann. Statist.*, 35(6):2450–2473, 2007.

[Yao88] Yi-Ching Yao. Estimating the number of change-points via Schwarz' criterion. *Statist. Probab. Lett.*, 6(3):181–189, 1988.